Classification Influence Index and its Application for *k*-Nearest Neighbor Classifier

Sejong Oh

***Abstract*—**Classification is an important topic in machine learning and bioinformatics. Many datasets have been introduced for classification tasks. A dataset contains multiple features, and the quality of features influences the classification accuracy of the dataset. The power of classification for each feature differs. In this study, we suggest the Classification Influence Index (CII) as an indicator of classification power for each feature. CII enables evaluation of the features in a dataset and improved classification accuracy by transformation of the dataset. By conducting experiments using CII and the k-nearest neighbor classifier to analyze real datasets, we confirmed that the proposed index provided meaningful improvement of the classification accuracy.

***Keywords*—**accuracy, classification, dataset, data preprocessing

I. Introduction

CLASSIFICATION is the division of samples in a dataset to specific classes. Classification is currently a popular topic in the field of machine learning because it can be applied to wide areas of research such as pattern recognition, image processing, and document classification. Classification is also used in bioinformatics, such as protein function prediction and microarray data classification. The ultimate goal of classification tasks is to improve the classification accuracy. Classification accuracy is highly dependent on the quality of the dataset. Each feature in a dataset contains data value to classify samples in the dataset. Therefore, finding good features for classification is important during classification analysis. For example, 'sex chromosome' is a better feature than 'height and weight' for classification of male and female. If a dataset has many features, the contribution of each feature for accuracy differs. When new datasets are developed for some classification task, it may be necessary to evaluate each candidate feature and select highly qualified features. In previous studies, statistical tools such as the mean and standard deviation have been used to evaluate candidate features. However, these tools only evaluate a few statistical characteristics of a feature and do not directly express the degrees of classification power. In this study, we propose the Classification Influence Index (CII) as an indicator of classification power for individual features in a dataset. If the CII value of feature f_1 is greater than that of feature f_2, f_1 has a greater impact on classification accuracy than f_2. Furthermore, we can transform feature values according to their CII value, which is expected to improve the classification accuracy. The remainder of this paper is structured as follows. In section 2, we describe the related works of this study. Section 3 shows a formal description of CII with some related definitions. In section 4, the experimental results obtained using actual datasets are discussed. Finally, concluding remarks are presented.

Sejong Oh is with WCU Nanobiomedical science department, Dankook university in Korea (phone: 41-550-3484; fax: 41-550-3480; e-mail: sejongoh@ dankook.ac.kr).

II. Related Works

As mentioned above, the ultimate goal of classification tasks is to improve the classification accuracy. There are two approaches used to accomplish this task: the *data-oriented* approach and *classifier-oriented* approach (*see* Table 1). The proposed CII belongs to the data-oriented approach. The purpose of the data-oriented approach is to produce the best dataset for a specific classification task. A dataset contains multiple features, samples, and class labels for the samples. Feature selection, which is also known as variable selection or attribute selection, is the selection of relevant features from a high number of candidate features. If a dataset contains many features, a long training/testing time is required for classification. Furthermore, many features do not guarantee high classification accuracy; therefore, it is necessary to remove useless features from candidate features and select only appropriate features. Feature selection techniques are grouped by filter, wrapper, and embedded methods [1]. Recently, F-test [2] and ANOVA [3] were introduced as filter methods. Ressom *et al*. suggested a hybrid ACO-SVM algorithm and applied it to select eight features from 228 candidate features [4]. Sun and Wu proposed a new feature selection algorithm based on the traditional Relief technique [5]. FSDD [14] and MRMR [15] are also one of new feature selection algorithms. Cui *et al*. suggested a new ranking and selection method for microarray experiments [6]. Several groups have also attempted to reduce instances (data samples) from training data [7-9], which reduces the time required for the training step and can improve the classification accuracy. After selecting features and instances, a filtering and normalization step known as preprocessing in which noise instances are removed is conducted [10]. There are several types of normalization, centering normalization, scaling normalization, and intensity dependant normalization. The purpose of normalization is to improve the efficiency of data analysis and calculation during training/testing tasks [11-13].

The *classifier-oriented approach* is another method of improving classification accuracy. This approach is designed to improve known classification algorithms by distance function modification, adoption of fuzzy theory, use of kernel functions, and so on. *k*-nearest neighbor (KNN), artificial neural network (ANN), and support vector machine (SVM) are well-known classification algorithms.

We recently proposed the use of the *R*-value to evaluate a dataset [14]. This proposed method is based on the ratio of overlapping areas among classes in a dataset. A high *R*-value

for a dataset indicates that it contains wide overlapping areas among its classes, and indicates that the classification accuracy of the dataset may become low. This supports the notion of the 3-level degree of overlap: whole dataset level, single class level, and two class level. The proposed CII value focuses on each feature, whereas the *R*-value focuses on each class and whole dataset. In the next section, we describe the details of CII.

III. Classification Influence Index (CII) and its Application

The classification influence index (CII) for a feature f_i expresses the contribution level for classification accuracy of a dataset that f_i belongs to. The CII can be calculated from the difference between classification accuracy of the whole dataset and a subset that excludes f_i. If the CII value of f_i is high, the classification accuracy of the entire dataset will likely decrease if we remove feature f_i from the entire dataset. Therefore, the CII value shows the influence of f_i in a given dataset. If the CII value of f_i is negative, the classification accuracy of the entire dataset will increase if we remove feature f_i. In this case, feature f_i has a negative influence on the classification accuracy. We can expect improvement of the accuracy in response to removal of features that have negative CII values. We provide a formal definition of the CII value after defining some notations.

TABLE I
Improvement of Classification Accuracy Using the Data-Oriented and Classifier-Orient Approaches

Approach	How classification accuracy is improved
Data-oriented	Feature selection Reducing number of instances, Noise filtering Normalization
Classifier-oriented	Develop new distance metric Merge fuzzy concept to classifiers Make hybrid-classifier (ex. KNN-SVM) Merge kernel method to classifiers Merge the above methods together

Let's suppose $D = \{f_1, f_2, f_3, .. , f_n\}$ is a n-dimensional dataset where f_i is a i-th feature of D

- S_j : j-th sample of D
- $S_j(f_i)$: feature value of f_i in sample S_j
- D_i : subset of D defined by $D - \{ f_i \}$
- $CA(X)$: classification accuracy for given dataset X

Definition 1. Absolute classification power of f_i denoted by $ACP(f_i)$ is the difference between classification accuracy of the whole dataset D and subset $D - \{f_i\}$.

$$ACP(f_i) = CA(D) - CA(D - \{f_i\}) \quad (1)$$

$ACP(f_i)$ expresses the influence of the degree of classification accuracy of f_i. In many cases, $ACP(f_i)$ has a very small value such as 0.0012 because the maximum value of accuracy is 1. Furthermore, if a dataset has many features, the influence of each feature become weak, whereas the influence of each feature in a small feature dataset remains strong. Therefore, we devise CII as a relative degree of influence between features.

Definition 2. The maximum value of *Diff*() denoted by max_D is the maximum value in $\{ ACP(f_1), ACP(f_2), ACP(f_3), .., ACP(f_n) \}$. If all $ACP(f_i)$ values are negative, max_D is the absolute value of the minimum value in $\{ ACP(f_1), ACP(f_2), ACP(f_3), .., ACP(f_n) \}$.

Definition 3. Classification Influence Index of feature f_i denoted by $CII(f_i)$ is given by:

$$CII(f_i) = Diff(f_i) / max_D \quad (2)$$

As we can see, the maximum value of $CII(f_i)$ is 1 and the minimum value of $CII(f_i)$ is not fixed. The classification accuracy of CII can be calculated for any types of classifiers including KNN, ANN, and SVM. As shown above, $CII(f_i)$ depends on the classifier. Accordingly, it is necessary to select a classifier to test CII. In this study, we used the KNN algorithm for CII because it is simple but strong for classification tasks.

We now introduce CII transformation that is for improvement of classification accuracy. In this approach, we modify the original dataset based on CII value.

Definition 4. CII transformation is a process of data transformation by the following steps:

Step 1. Remove feature f_i from a given dataset that has a negative $CII(f_i)$
Step2. Transform the rest of the feature values using $CII(f_i)$. The feature value $S_j(f_i)$ is transformed into $S_j(f_i)'$ by equation (3).

$$S_j(f_i)' = S_j(f_i) \times (1 + CII(f_i) \times w)$$
where, w is a weight value that controls the influence of $CII(f_i)$ and $0 \le w$ (3)

From the equation, we can see that if $CII(f_i) > CII(f_k)$, then the variation of $S_j(f_i)'$ is higher than $S_j(f_k)'$. This difference in variation between features leads to improved classification accuracy. The reason for this improvement is discussed in Section 4.2.

IV. Experiment and Discussion

A. Result of Experiments

To test the effects of class-space reduction, we selected six real datasets from the UCI machine learning repository (http://archive.ics.uci.edu/ml/). Table 2 summarizes these datasets. The KNN classifier was tested on the reduced datasets. The number of nearest neighbors was 7. To ensure the credibility of the classifications, we use the k-fold test where k= 4.

Figure 1 shows the examples that were subject to CII evaluation. The liver dataset was used to classify liver disorder patients. The first five features were all blood tests thought to be sensitive to liver disorders that might arise from excessive alcohol consumption. As shown in Figure 1, features *f1* and *f2* exert a negative influence on the classification accuracy, whereas *f3* and *f5* have a strong positive influence. These findings indicate that the *mcv* and *alkphos* features should not be used for liver disorder classification. Medical scientists may develop indicator materials using CII analysis for diagnosis of liver disorder patients. In the case of the balance scale dataset, no features exert a negative influence on classification. However, features *f3* and *f4* are better than *f1* and *f2*. CII values express the relative degree of influence of features in a dataset. The CMC dataset has many negative features based on negative CII values. To investigate the relationship between the degree of influence and classification accuracy, the use of ACP (absolute classification power) is more useful than the CII.

TABLE II
SUMMARY OF DATASETS

Dataset	# of samples	# of classes	# of features
Liver	345	2	6
Wine	178	3	13
Soybean	307	14	35
Parkinson	195	2	22
Balance scale	625	3	4
CMC	1473	3	9
Hayes roth	132	3	5
Pima Indians diabetes	768	2	8

To show the usefulness of the CII value, we proposed CII transformation. We removed features that had negative CII values from the given dataset, and then transformed the values of the remaining features according to their CII value. Figure 2 in appendix shows the results of the experiment. In each dataset, we tested the classification accuracy of the original dataset (Original), sub datasets of the original dataset that had no features with negative CII value values (RM), and six transformed sub datasets based on the weight value *w* in Equation (3) (*w*0.2, *w*0.3, *w*0.4, …). The data presented in Figure 2 indicated the following:

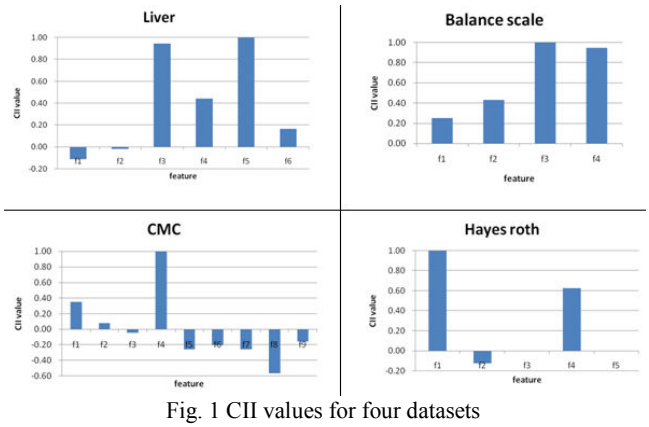


Fig. 1 CII values for four datasets

- Each graph showed improved classification accuracy of the original dataset when negative features were removed and the remaining features were transformed according to the CII values. These findings indicate that the CII transformation is useful for improving the classification accuracy.
- The influence of removing negative features was greater than that of transforming feature values for improving the classification accuracy.
- The weight value *w* influences the improvement of accuracy. In most cases, *w* was < 1.0, which produces the best classification accuracy. In some cases, such as the soybean dataset, *w* was > 1.0.
- Some datasets such as the balance scale dataset have no negative features; therefore, they are not improved in RM dataset.
- If CII values of features show little difference, then there is little improvement of accuracy by transforming the feature values (*see* Parkinson and CMC).
- The accuracy of some datasets was not improved when compared with the original dataset (*see* Pima Indians diabetes). The CII transformation only works well to improve the accuracy of a dataset if it has a large number of negative features and variation of CII values of features is high.

Table III summarizes the number of reduced features (negative features) and improvement of classification accuracy for eight benchmarking datasets. The soybean and CMC datasets had a high ratio of negative features. The improvement of accuracy was between 4% and 15% except for Pima Indians diabetes.

Figure 3 in appendix shows comparison of CII transformation to FSDD, ReliefF, and MRMR. We compare classification accuracy of feature selected datasets derived by the feature selection algorithms and proposed CII transformation. The number of selected feature for each dataset is (# of original features - # of negative features) in Table 3. In every graph, the feature selected dataset derived by CII transformation shows best accuracy, and it means that CII value is better feature evaluation indicator than evaluation functions of other feature selection algorithms.

B. Discussion

The CII value expresses the influence of each feature for classification accuracy. This value is used for development of a new dataset for specific classification tasks. When a dataset is prepared, we gather data from several types of experiments. Each experiment then becomes a feature in a dataset. When conducting an experiment, the degree of the contribution of each feature is not known. After gathering candidate feature data (experiment data), we can measure the quality of each feature by evaluating the CII value. We then remove the negative features and identify the best *w* value for transforming the data. Finally, we select the optimal dataset for the

TABLE III
REDUCED FEATURES AND IMPROVEMENT OF ACCURACY BY CII VALUE

Dataset	# of original features	# of negative features	Improvement of accuracy
Liver	6	2	4%
Wine	13	2	6%
Soybean	35	18	15%
Parkinson	22	3	11%
Balance scale	4	0	5%
CMC	9	6	4%
Hayes roth	5	1	5%
Pima Indians diabetes	8	1	0%

classification task. The CII value also can be used for feature selection. High ranked features according to the CII value can be selected for the new dataset.

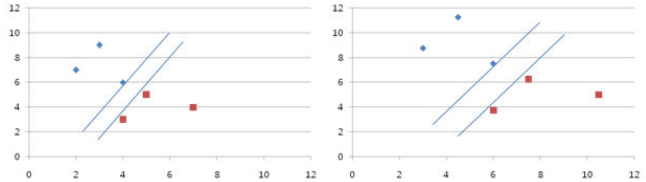
Transforming feature values according to the CII value leads to improved classification accuracy. The effect of CII transformation is reinforcement of the influence of features according to the CII value. Each feature has a different CII value. If a feature has a high CII value, its variations are highly enlarged after transformation. These results in changes in the location of data points in a dataset as well as changes in the classification power (Figure 4). Figure 4(a) is a dataset before transformation and Figure 4(b) is a dataset after transformation according to:

$$S_f(f_x)' = S_f(f_x) \times (1+ 0.5 \times 0.5)$$
$$S_f(f_y)' = S_f(f_y) \times (1+ 1 \times 0.5.$$

The boundary between $class_1$ and $class_2$ becomes more distant with transformation, which leads to improved classification accuracy

The weak point of the proposed CII approach is a high time complexity. To obtain a CII value for each feature, it is necessary to run n times classification work, where n is the number of features. To obtain the optimal w value, 6~10 extra classification works are required. If a dataset contains a large amount of sample data, CII take a long time to calculate. In such a case, we can randomly select the proper number of samples for the CII task. In the case of datasets in table 2, the runtime for determination of the CII value and identifying the optimal w value is 0.5 – 5 seconds, which is reasonable for practical use

One of the factors that influences classification accuracy is correlation between features. For example, in the male/female classification task, independent height and weight have a low relationship with classification accuracy, but combined they have a strong relationship with classification accuracy. In this case it is difficult to measure the correlation, and background knowledge for the given features is sometimes required.



(a) before transforming (b) after transforming

Fig. 4 Effect of transforming feature data

FSDD, ReliefF, and MRMR are newest feature selection algorithms and have been shown good performance for huge features datasets such as microarray, but they are not good in our classification test. The reason is that the datasets in our experiments have small numbers of features than microarray. In the case of microarray, it has thousands of features and contains many useful/useless features. If an algorithm is missing to select a useful feature, the influence of it is restrictive. However, if a given dataset has small numbers of features, missing a useful feature may bring critical decrease of classification accuracy. Table IV shows miss selection of previous feature selection algorithms. Therefore, CII value is powerful to make feature selected datasets than previous feature selection algorithms in the small features dataset.

V. CONCLUSION

There were various approaches to improve the classification accuracy of the target area. Classification accuracy is highly dependent on the quality of the dataset, more specifically, the quality of features in the dataset. Here, we suggest the CII value as an evaluation measure for each feature. We confirm that this index is useful for development of a high quality dataset based on experimental results. The CII value can be used for any type of classifier, and the measuring power is dependent on the characteristics of the base classifier. The CII value assumes that each feature is an independent variable, and it cannot capture correlation between features. It is known that correlation influences classification accuracy. If we combine CII and measure the correlation, it may lead to improved accuracy. This will be addressed in future studies.

TABLE IV
EXAMPLE OF MISS SELECTIONS FOR LIVER DATASET

	f1	f2	f3	f4	f5	f6
CII value	-0.11	-0.02	0.94	0.44	1.0	0.17
CII selects	X	X	O	O	O	O
FSDD selects	O	X	O	O	X	O
ReliefF selects	O	X	O	O	X	O
MRMR selects	O	X	O	X	O	O

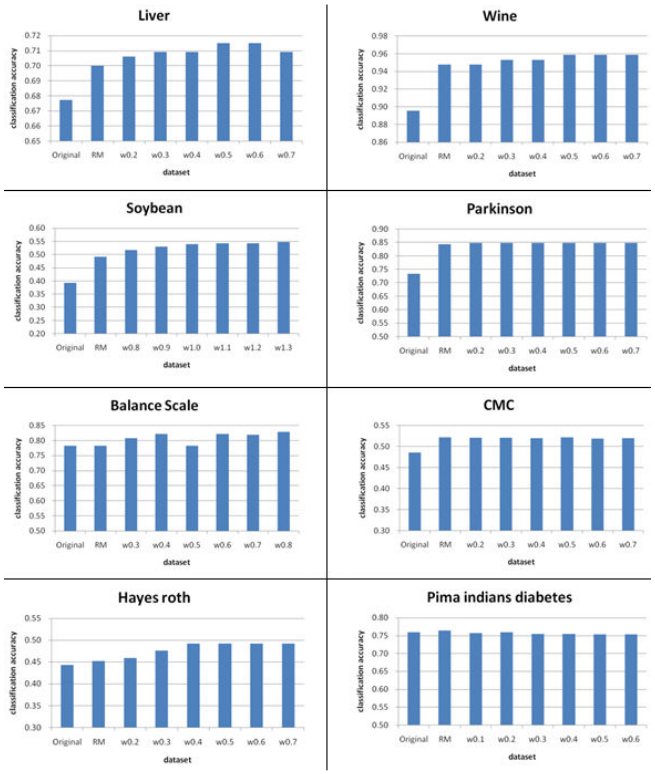


Fig. 2 CII Variations in classification accuracy according to class-space reduction

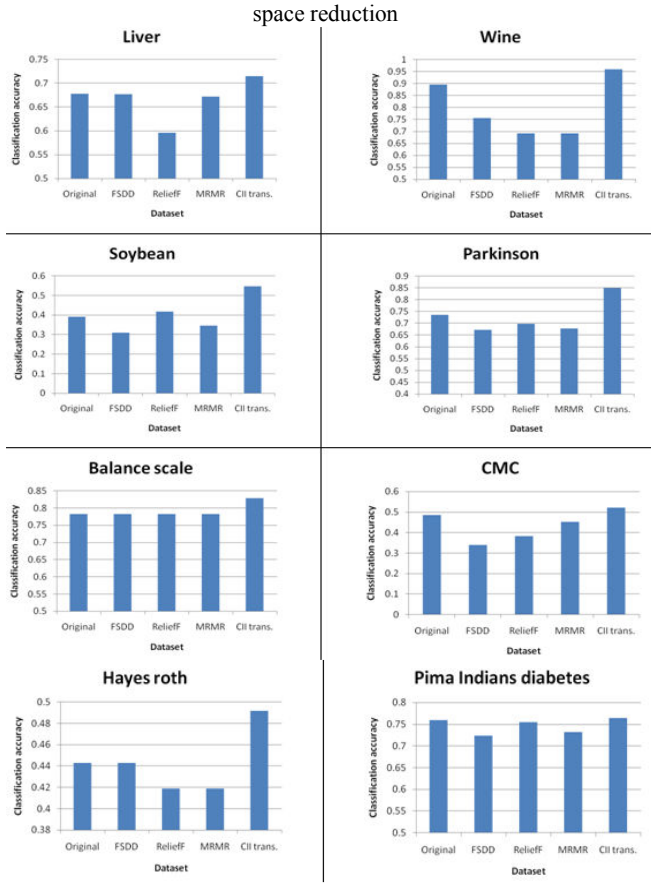


Fig. 3 Comparison of proposed method to FSDD, ReliefF, and MRMR

ACKNOWLEDGMENT

This study was supported by grant No. R31-2008-000-10069-0 from the World Class University (WCU) project of the Ministry of Education, Science & Technology (MEST) and the Korea Science and Engineering Foundation (KOSEF).

REFERENCES

[1] Y. Saeys, I. Inza, and P. Larranaga, A review of feature selection techniques in bioinformatics, *Bioinformatics*, 23, 2007, pp.2507-2517.
[2] G. Bhanot, G. Alexe, and B. Venkataraghavan, A robust meta classification strategy for cancer detection from MS data, *Proteomics*, 6, 2006, pp.592–604.
[3] P. Jafari, and F. Azuaje, An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors, *BMC Med Inform Decis Mak*, 6, 2006, p. 27.
[4] H. W. Ressom, R. S. Varghese, and S. K. Drake, Peak selection from MALDI-TOF mass spectra using ant colony optimization, *Bioinformatics*, 23, 2007, pp. 619–626.
[5] Y. Sun, and D. Wu, A RELIEF Based Feature Extraction Algorithm, *in Proceedings of the 2008 SIAM International Conference on Data Mining*, 2008, pp.188-195.
[6] X. Cui, H. Zhao, and J. Wilson, Optimized Ranking and Selection Methods for Feature Selection with Application in Microarray Experiments, *J Biopha Stat*, 20, 2010, pp.223-239.
[7] S. Xu, Q. Luo, and H. Li, Time Series Classification Based on Attributes Weighted Sample Reducing KNN, *Proceedings of the 2009 Second International Symposium on Electronic Commerce and Security*, 2009, pp.194-199.
[8] Y. Liao, and X. Pan, A New Method of Training Sample Selection in Text Classification, *2010 Second International Workshop on Education Technology and Computer Science*, 2010, pp.211-214.
[9] Y. Xu, L. Zhen, and L. Yang, Classification Algorithm Based on Feature Selection and Samples Selection, *Lecture Notes in Computer Science*, 5552, 2009, pp.631-638.
[10] B. T. McBride, and G. L. Peterson, Blind Data Classification using Hyper-Dimensional Convex Polytopes, *Proceedings of 17th International FLAIRS conference*, 2004, pp.1-6.
[11] J. Schuchhardt, D. Beule, and A. Malik, Normalization strategies for cDNA microarrays, *Nucleic Acids Research*, 28, 2000, E47-e47.
[12] W. Wu, E. P. Xing, and Connie Myers, Evaluation of normalization methods for cDNA micro-array data by k-NN classification, *BMC Bioinformatics*, 6, 2005, p.191.
[13] G. Collewet, M. Strzelecki, and F. Mariette, Influence of MRI acquisition protocols and image intensity normalization methods on texture classification, *Magnetic Resonance Imaging*, 22, 2004, pp.81-91.
[14] S. Oh, A New Feature Evaluation Method Based on Category Overlap, *Computers in Biology and Medicine*, 41, 2011, pp.115-122.
[15] J. Liang, S. Yang, A. Winstanley, Invariant optimal feature selection: A distance discriminant and feature ranking based solution, *Pattern Recogn.*, 41, 2008, pp.1429-1439.
[16] C. Ding, H. Peng, Minimum Redundancy Feature Selection from Microarray Gene Expression Data, *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, 2003, p.523.

Sejong Oh received a Doctor, Master, and Bachelor degree in Computer Science from Sogang University, Korea, in 2001, 1991, and 1989. From 2001 to 2003, he was a postdoctoral fellow in the laboratory for Information Security Technology at George Mason University, USA. Since 2003 he joined the Department of Computer Science at Dankook University, Korea, and is currently associate professor in WCU Research Center of NanoBioMedical Science. His main research interests are bioinformatics, information system, and information system security.