International Journal of Engineering & Technology, 7 (4.31) (2018) 272-279

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Privacy Preserving Technique for Mitigating Anonymity Attack in Pervasive Social Networking Applications

Nur'Ayuni binti Adnan[1], Manmeet Mahinderjit Singh[2]*, Aman Jantan[3]

School of Computer Sciences
University of Science Malaysia
Pulau Pinang, Malaysia
**Corresponding author E-mail: manmeet@usm.my*

Abstract

Pervasive Social Networking (PSN) applications become more popular in the last few years. The uses of PSN applications through mobile devices such as smartphones, tablets will lead to the security and privacy issues. This is because users tend to share their personal information with the third party organizations such as applications in mobile devices. Due to the development of social network, the security and privacy need to be improved as well as others to make sure that all the user's information is protected securely in social network (SN). In this study, we will focus more on the privacy issues on how to preserve the privacy of user's data from being known by the third party. The dataset of PSN application will be tested using data mining tool, which is Weka, in order to identify the optimal technique and classifier that can be applied to conceal the information. Then, a new enhanced base learner will be proposed, which is masking technique algorithms will be implemented into the dataset of PSN application at the end of this research.

Keywords: *Pervasive Social Networking (PSN); Privacy preserving technique; Social Network (SN)*

1. Introduction

PSN is the combination of Pervasive Networking and Social Networking [1]. Pervasive is also known as ubiquitous which means existing everywhere, while social networking is a medium for users to communicate to one another through online services. This gives the meaning of PSN as social networking that connected the users through mobile devices anytime and anywhere. PSN is also known as Mobile Social Networking (MSN) [2].

Recent advances in technology lead the uses of mobile devices such as smartphones and tablets become more popular. The rapid development of mobile technology combined with improvement in communication capabilities enables people to access to technology wherever they may be, at all time. According to the survey conducted by Warren [3], concluded that the number of mobile subscribers accessing MSN such as Facebook and Twitter increased by 112% and 347% from January 2009 to January 2010. This is because some of the improvement has been made in order to make it interesting services and ways of engaging in social interaction through mobile devices. Based on the survey conducted, people nowadays prefer to use mobile devices instead of a personal computer (PC) to access to the online social networking (OSN).

One of the important issues of PSN applications is security and privacy of sensitive information [4]. Due to the development of SN, most of the PSN applications indirectly required users to become a system and policy administrators in order to protect their contents online. Sensitive information such as financial and medical information, both of information could harm to users if the third party get the information. There are several of security and privacy attacks that can happen to PSN applications such as anonymity attack, linkability attack and so on.

We are conducting this research based on the problem statements provided by the previous researchers. In [5], said that PSN applications need the user allow access to his/her social network profile information and it will compromised with user identity and as for [6], said that anonymity attack occurred when information of the user is linked with publicly database even the data reveal is anonymous dataset. From the problem statements, we can see that even the data being conceal securely, there still chances for the third party to steal the information. As for this research, we will actually explore various privacy preserving data mining techniques in PSN application and enhance the base privacy technique by adding masking algorithm to tackle anonymity attack from occurred.

In this research, we will focus on privacy issues in PSN, on how to protect the user's privacy from being known by the third party. We will work out with the dataset that we are collected using PSN application, which is NFC Florist application [7]. Then, experiments will be conducting to identify which technique and classifier that is suitable to be a base technique in order to preserved privacy of data mining using Weka and we will also focusing on data mining approach only.

2. Literature Review

2.1. Pervasive Social Networking

The basic features of PSN includes mobility, content, accessible, connect, and also context. Mobility is defined as the ability to move or be moved freely and easily, such people nowadays prefer using mobile devices instead of PC to do workload because it is more convenient compared to the PC. Content means that the user can get the information easily which this will leads to the user to know anything at anytime. Accessible is one of the features that can make the user to access to the Internet anytime and anywhere because it is always on and available. For connecting, the mobile device need to be connected to the Internet in order to used PSN applications. Lastly is context, where it is the set of circumstances that surround a situation or an event which this features can used to describe the situation such as user's identity, location and also activity of user.

PSN applications served based on the three services which are location-aware, proximity-aware and also centralized servers [8]. Location-awareness is the one that provides the location of its user from their GPS-enable mobile devices, using the information to locate nearby friends, provide recommendations, and allow user to discover their surroundings. The applications that serve location-awareness are FourSquare, SocialAware, WhozThat and also Serendipity. While, proximity-aware is the one that allows its user to use Bluetooth-enable or Wi-Fi connectivity to find and communicate nearby friends and others with similar profiles and interest. The applications that involved in proximity-aware are FaceTime, Jambo, BlueDating and Toothing. For centralized server, it is a remote server, which mobile user interaction to get the services including exchanging messages, viewing profiles, downloading and playing games, streaming video and also making professional contact. The applications that involved in centralized server are Facebook, Twitter, Friendster, YouTube and LinkedIn.

Due to the development of technologies nowadays, many challenges that can be faced by the developers in order to developed more PSN applications. One of the challenges that can happen is performance. This challenge may happen when the user is expected that the performance of the services in the mobile devices is the same level as performance that they enjoy through the desktop. Other than that, the challenge that may happen is communication, where future wireless technologies may someday support more efficient opportunistic communications. For personalization aspect, it will tend the users to requesting services with friendly interfaces and the ability to match profiles, backgrounds, and contexts. Next PSN challenge include friend discovery, where it will supporting dynamic changes in context and benefiting from historical information. Last but not least is security and privacy, where it will prevent data misuse or breaches of confidentiality. In this research, we will focus only on security and privacy challenges that happened in PSN applications.

2.2. Security and Privacy Issues in PSN

There are many issues regarding to the security and privacy of PSN applications, which need to be focused more by the developers. From the previous researchers, there are many enhancements that have been deployed, but then, it still cannot fully protect the security and privacy of the user. Table 1 shows some of the PSN applications with the problems and proposed solutions by previous researchers.

2.3. Privacy Attack - Anonymity Attack

Based on the research done by [6], she said that anonymity attack is an attack that can occurred when information of the user is linked with the publicly database even the data is reveal is anonymous dataset. While as the research done by [5], he said that anonymity attack is an attack that happen when the OSN of the user is associates with the system and it will compromised with the user's identity.

Table 1 : Existing PSN Applications with the Problem and Proposed Solution

Application	Problem	Proposed
MobiClique and Looptmix [11] Function: It connect the users using process matchmaking.	In the existing matchmaking, there are two ways of implementation that will lead to privacy issues; When device reveal the owner's profile information in public using Bluetooth method. When there is involvement of trusted server for matchmaking operation where all the information of the user will be stored in the server.	Privacy-preserving Matchmaking Protocol. It consists of two principles where; Ensures that user's privacy will not reveal unnecessary private information to other user. Will not use any trusted server in order to do the matchmaking operation.
Google Latitude [2] Function: It connect the users using process matchmaking.	Typically, most of the PSN application will be limiting to the PSN vision which does not complement virtual interactions with the physical one.	Social networking middleware services. It combines both physically and socially related to find each other and perform activities of common interest.
Jambo [12] Function: Connect the users using location-based services, where others can detect the location of the user using GPS application	It enables the users by relaying on location information and coordination. This will lead the user to locate and meet friends in nearby location. This service will lead to the privacy issues where third-party can get the location of the user.	There are three approaches introduced; Identity Servers and Anonymous Identifier (AIDs) Virtual Individual Servers Re-socializing Social Network

Anonymity attacks basically divided by two types which are direct anonymity and indirect anonymity. The direct anonymity attack is an attack that will compromise with the user's anonymity to get the information and also the location of the user. While for indirect attack, it will take place when pieces of information will be mapped back to the user's identity to get the information. Table 2 shows that the previous researches that have been done on anonymi1ty attack.

Table 2 : Previous Researches on Anonymity

Proposed Solution	Advantage	Disadvantage
Multidimensional suppression [13] which involved two stages: Anonymization method Suppression method	Multidimensional suppression is better than single-dimensional suppression because it suppresses a certain value in all tuples without considering values of attributes.	Only use C4.5 decision tree.
Classification Tree-Based K-anonymity using Masking operations (CTKAM) [14].	Masking of user identity done by using privacy technique.	Only use C4.5 decision tree.
Two methods used are generalization and	Clustering technique used in Weka.	Only focused on anonymization

suppression and it will evaluated using K-Means, Expectation Maximization and Density [15].		technique
The researches conduct a survey (collected the data) and also interviewing to investigate how many participants that are known about the important of their own privacy [16].	Merely survey on important of privacy.	Not implemented.
Weka-GDPM (Geographic Data Preprocessing Module), which integrates geographic databases and the classical data mining toolkit Weka in the field of spatial data mining [17].	Automatically generates data at two granularity levels without using prior knowledge and provide support for both distance and topological spatial relationship.	Only focus on geographical information.
Weka-STPM (Semantic Trajectories Preprocessing Module) which can preprocess from trajectories to the semantic trajectories [18].	Focus on moving objects.	Focus only on geographical information.

2.4. Data Mining

With the advanced of technology, nowadays, there are abundance of information about individuals that one can be obtained within seconds. This information could be obtained through mining or just from information retrieval. For instance, in health areas, in their system, there are information of the patients that have been stored for almost a decade and even more than that. This data is called as data mining. Data mining is the process of extracting information from the big data. The main purpose of data mining is to find the patterns and also relationship of the dataset. In order to get the information through the dataset, a tool needed to support to get the information. There are many tools that provided in order to extract the data mining such as Weka, RapidMiner, KNIME and so on.

2.5. Privacy Preserving Data Mining

According to [9], there are many approaches that have been adopted for privacy preserving data mining. The one that have been classified is as in Figure 1. There are five dimensions that involved in privacy preserving data mining, which are data distribution, data modification, data mining algorithms, data/rule hiding and also privacy preservation.

Data Modification: It is used in order to change the original values to unique values of the database to ensure high privacy protection. Techniques that can be used such perturbation, randomization, anonymization, swapping, and also sampling.

Data Mining Algorithm: The phase that data modification will take place. The classifications that involved in this dimension are decision tree, association rule mining, clustering algorithm, rough sets and also Bayesian networks.

Data/Rule Hiding: Dimension that is used to hide whether raw data or grouped data [10]. Data hiding means the process of hiding the sensitivity of data values such as names, phone numbers, address, and so on. While rule hiding is the process of protecting the confidentiality knowledge in data, such as association rule.

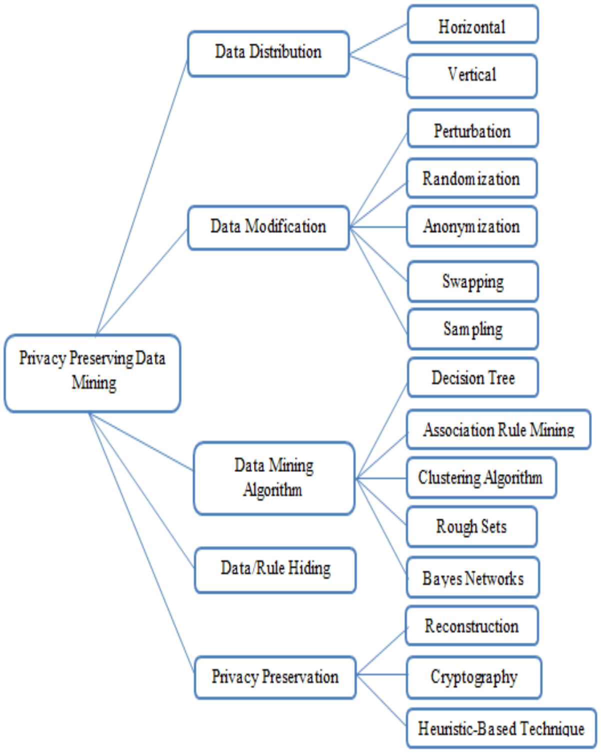


Fig. 1: Taxanomy of Privacy Preserving Data Mining [9]

Privacy Preservation: It is used for the modification of the selective data. This is important in order to achieve higher utility for the modified data given so that the privacy is not affected. There are three techniques that applied to this dimension which are reconstruction, cryptography and heuristic-based technique.

2.6. Data Modification

Data modification is used in order to change the original values to unique values of the database that need to be allowed to the public and this way to ensure high privacy protection. Techniques that will be used in data modifications such as perturbation, randomization, anonymization, swapping and also sampling.

Pertubation: The process of changing the original data with some synthetic data values so that the data is not differ to the original data [10]. Usually in this technique, random noise will be added to data and make sure that the random noise keep the signal from the data so that the pattern of the data still can be estimated.

Randomization: The process of twisted the data in such way that the values of the data cannot be determined which one is the correct information and which one is the incorrect information [10]. This technique basically will used a technique by adding the noise to the original data.

Anonymization: The process of hidden the identity or sensitive data of the user in the database. For this technique, usually the explicit data such as names will be removed instead of encrypted the names, but still it can lead to the linking attack if the quasi identifier is still in the database such as age, country, zip code and so on.

Swapping: The process where the values of the neighboring records are swapped. Unlike randomization which is implemented on independent data. By using this technique, the values of records are interchanged and the original data will not be revealed to the researchers.

Sampling: The process of releasing a small portion of data in random [10], and then it will be the input for an algorithm. It is statically dependable. This technique will make the analyst to spend more time in models, but it will take less time for modelling the results.

Table 3 shows that the advantages and disadvantages of data modification techniques based on the previous researches.

Table 3 : Advantages and Disadvantages of Data Modification Techniques

Techniques	Advantages	Disadvantages
Perturbation [10]	In this technique different attributes are preserved independently.	This technique will cause loss of information and the original data values cannot be regenerated.
Randomization [10]	This is the simple method in order to hide the information of user. Better efficiency compare to cryptography based PPDM technique [20].	This method will not effect for multiple attribute databases and will cause loss of the information of the user.
Anonymization [10]	Identity or sensitive data about record owners are to be hidden.	This technique will cause linking attack and heavy loss of information.
Swapping [19]	Simple procedure and can performed on all potential key variables.	May take time in order to do the swapping.
Sampling	It is efficient if the data is in a small group.	If the data is in a big group of data, it may expose the valuable of the data.

3. Implementation

In order to achieve the aim of this research, we are conducting two phases to complete the task. Where for the first phase, we are conducting experiments using data mining tool which is Weka, in order to choose the base privacy technique that can be used to conceal the data. For the second phase, we are going to implement the masking technique algorithms that is propose by Nagmode [14] to ensure the privacy of the user's information more secure.

For the dataset, we will collect the dataset using PSN application which is NFC Florist application[7]. Basically this application will need the user to store their own information in the application. The application also can stored the location of the user when the user comment on the SN through the application. The dataset has 2238 instances with 18 attributes. The class that we have chosen for this dataset is 'Like' and 'NoLike'. The reason we are choosing that classes is because regarding the because regarding to the research done by the researchers from University of Cambridge in UK, the researchers found that they can get the personal information of the user using Facebook Likes alone. From the Facebook Like, they actually can predict accurately the information of the user such as race, age, personality and so on.

From the experiment conducting, we also will provide the performance evaluation criteria such as confusion matrix including accuracy, True Positive Rate (TPR), False Positive Rate (FPR) and also the Forecast Error Measurement such KAPPA Statistic, ROC Area, Mean Absolute Error (MAE) and so on.

3.1. Phase 1

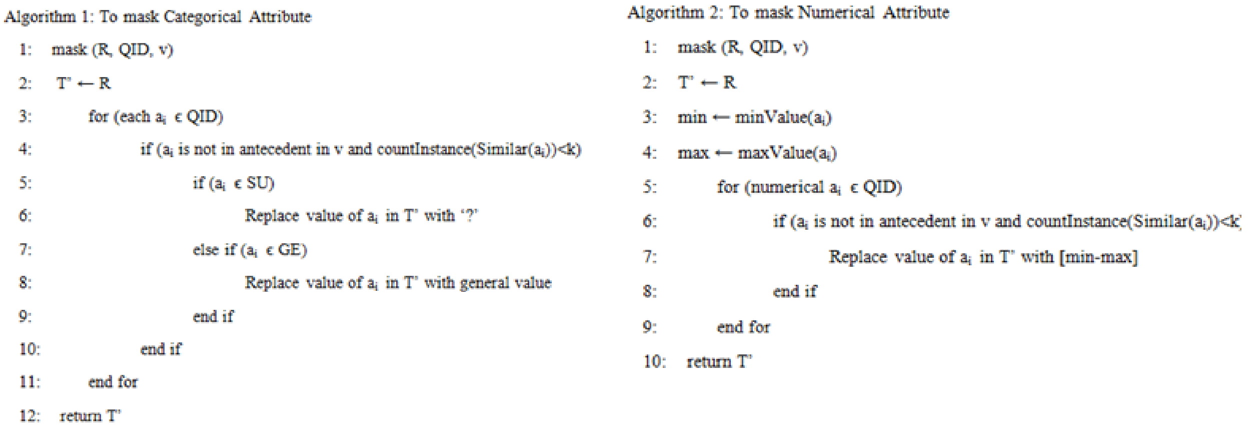
Before the experiments are conducted, we are tested the dataset with the feature selection such as CfsSubset Evaluation and Best-First, Correlation Attribute Evaluation and Ranker and also Info Gain Ratio Attribute and Ranker. In order to obtain the best result of accuracy from the database, performing feature selection is a must where only certain attributes that is important needed to be tested without any noise. After performed the feature selection, the dataset will be experimenting using certain techniques such as AddNoise, Randomize, Datafly, SwapValue and also Resample. For the classifier, we are using IBk, Naïve Bayes, J48 and also Random Tree.

In this research, we are going to combine the technique with the classifier in order to get the optimal result. Based on the Table 4, it shown that the previous researchers that using the combination techniques to conceal the data more securely.

Table 4 : Table of Weka Techniques Used by Researches [20]

Techniques	Enhanced/ Existing	Filter	Weka Techniques
Randomization	Existing	UN	Randomization+ Random Tree [21]
Anonymization	Existing	UN	Datafly+ SVM [22]
Perturbation	Enhanced	UN	Perturb+C4.5 modified
Swapping	Existing	UN	SwapValue+ J4.8 [23]
Sampling	Existing	UN	Resample+ J4.8 [23]

***UN - unsupervised**



```
Algorithm 1: To mask Categorical Attribute
mask (R, QID, v)
 T' ← R
    for (each $a_i$ ϵ QID)
          if ($a_i$ is not in antecedent in v and countInstance(Similar($a_i$))<k)
                if ($a_i$ ϵ SU)
                      Replace value of $a_i$ in T' with '?'
                else if ($a_i$ ϵ GE)
                      Replace value of $a_i$ in T' with general value
                end if
          end if
    end for
 return T'
```

```
Algorithm 2: To mask Numerical Attribute
mask (R, QID, v)
T' ← R
min ← minValue($a_i$)
max ← maxValue($a_i$)
    for (numerical $a_i$ ϵ QID)
          if ($a_i$ is not in antecedent in v and countInstance(Similar($a_i$))<k)
                Replace value of $a_i$ in T' with [min-max]
          end if
    end for
 return T'
```

Fig 2 : Algorithms of Masking Techniques on Categorical and Numerical Attribute

3.2. Phase 2

According to the research done by [14], in order to preserve the privacy of the information, the researcher proposed Classification Tree-Based k-anonymity using Masking Operations (CTKAM) where it is efficient in handling both generalization and suppression techniques. The technique that we will implement in this research is masking techniques where both generalization and suppression will be used in order to preserve the privacy of the user's information. Both generalization and suppression will be performed only on quasi-identifiers where, it is a set of attributes that can be easily identify individual's information such as age, gender and so on.

Generalization is the technique where it will substitute the value with semantically similar value but less detailed value. Example of generalization technique is when there is a data that assign as BirthDate, it will replace the BirthDate to the BirthYear, where the data is still there but in general formed.

While for suppression, it is a technique where it will replaced the attribute value with a null value or specific symbols. For example, for the zipcode attribute, from 12345, it will replaced the zipcode to the 1234?. This is how the suppression technique worked.

Masking technique algorithms basically consists of two algorithms which are algorithm to mask on the categorical attribute and also numerical attribute. Both algorithms are shown as in Figure 2. The elements that involved in the algorithm such as follows:

- R: new anonymous dataset that will be assigned from T'
- QID: quasi-identifier such as age, gender, date, time and so on
- v: number of leaf node get from the height of the classification tree
- T': anonymous dataset
- a_i : attribute involved

- k : integer value
- SU: attribute set of suppression
- GE: attribute set of generalization

4. Result And EvaluationPhase 1

The objective on the Phase 1 is to determine the optimal technique and privacy in order to conceal the data from being known by the third party. Before the experiments conducted, the dataset of NFC Florist application will be performed feature selection to determine which attributes that is suitable and important for the dataset.

Based on the Figure 3, we can see that the results of accuracy on different classifier using different feature selections. From the result obtained for all the experiments conducted, the feature selection that shows the highest accuracy is Info Gain Ratio Attribute and Ranker with the help of Datafly technique and also J48 classifier which is 100% of accuracy. For the worst accuracy goes to the CfsSubset Evaluation and Best-First with the help of the AddNoise technique and Random Tree classifier which is only 73.45% of accuracy. So, from the results obtained, we decided to used Info Gain Ratio Attribute and Ranker because it gives the optimal result where the accuracy is higher when the dataset is applied with Datafly technique and J48 classifier.

4.2. Summary on Phase 1

As we can see from the experiment conducted, the result shows for all the selected attributes give the same result where the highest accuracy is on the J48 classifier with the help of Datafly filter that gives 100% accuracy.

From all the selected attributes methods used, we will focused on the Info Gain Ratio Attribute and Ranker method because from the results obtained, this method shows the synchronize pattern where among all the filters and classifiers used, it show the highest accuracy is on Datafly filter and J48 classfier. J48 achieved highest accuracy due to the way it classify the data efficiently and using information gain.

Based on the research done by Kantarcioglu (2009)[22], the researchers using anonymization technique also in order to preserved the privacy of the user data. The technique and classifier that the researchers used is Datafly filter with the help of SVM classifier. Why it is different with our research because in our research, we did not used SVM as one of the classifier tested in our data because of the attributes that involved in the NFC Florist dataset where there are some attributes that is used randomly hyperplan, which is the linear optimal that separate the decision boundary is randomly used. This can cause the result of poor accuracy.

4.3. Phase 2

The objective on the Phase 2 is to conceal user identity using masking technique algorithms. The base privacy technique that we obtained from the Phase 1, we will used that result to implemented the masking algorithms into the dataset.

As we can see in the Figure 4, it shown that the data collected before and after the masking technique algorithms are applied. From the figure, we can see that, the createdTime and createdDate which is in nominal attribute changes from the exact date and time to the more general date and time. As for the createdDate from 2/1/2017 change to the JANUARY only. This show that the attribute undergo generalization method where the data is changing from the exact data to general form. The same method goes to createdTime as well, from 15:00:30, it change the time to the PM.

While for numerical attribute, it will change the data into the range form. For example, as we can see from the Figure 4, the userID which is 3.0 is converted into the range form which is [3.0-13.0]. This range of number is comes from the lowest to the highest number that is stored in the system of the application. This is how the algorithm worked for masking the numerical attribute.

In order to test the accuracy of the dataset before and after the masking technique algorithms are applied, we run the experiments again using the same technique which is Datafly with different classifier. As we can see from the Figure 5, it shows the accuracy of the Datafly techniques without and with masking techniques. We can see that the accuracy of the IBk and Random Tree classifier is increasing while for Naïve Bayes and J48 show the same accuracy. This result can show the improvement of the accuracy after the masking technique algorithms is applied.

Based on the Table 5, it shown the performance evaluation for Datafly dataset with different classfiers before and after the masking technique algorithms is applied.

From there, as we can see, the results shows the improvement for most of the measurements. To be exact, the ROC area for both techniques basically has the improvement, where from only Naïve Bayes and J48 shows the ROC area is 1.000, as after the masking technique algorithms is applied, all the classifiers is given the ROC of 1.000. ROC area is the graph that is related to the TPR anf FPR. Where if the ROC area is approaching 1.000, it shows that the accuracy of the dataset is excellent, where there is 100% of True Positive (TP).

4.4. Summary of Phase 2

In order to secure the privacy of the user information, the masking techniques is applied in the NFC Florist dataset that we obtained from the Phase 1. Basically, from the experiment that we conducted in Phase 1, it give us the anonymous dataset where the dataset is already being conceal from being known by the third party and the linking data cannot be done. But, there are also chances for the adversary to get the linkage if the dataset is being conceal for only one method, this is why we implemented the masking techniques that is proposed by Nagmode, in order to preserved more securely the user's information [14].

As the masking techniques applied on the anonymous dataset, it will produced the enhancement of anonymous dataset. This dataset will provide more protection of the user's information from being linking to the publicly dataset. In masking techniques algorithm, we applied generalization and suppression method in order to preserve the privacy of the user. As the result, the enhancement anonymous dataset is being conducting again in the Weka in order to see the accuracy of the dataset whether the accuracy is being improved or it still in the same result.

After the experiment conducting in the Weka, it shows that the result of accuracy for the enhancement anonymous dataset is improved compared to the anonymous dataset. This result can be shown in Figure 5 where there are changes in accuracy of the dataset in the classifier of IBk and also Random Tree. These show that the masking techniques is performed well in order to preserve the privacy of the user information.

5. Discussions

5.1. Discussion on Phase 1

For the Phase 1, we already explored various techniques that can be used in order to tackle anonymity attack in the PSN application. We already conducted experiments in order to determine either the optimal privacy or base classifier that can be used in order to preserve the privacy of the user information.

From the literature review that we have done before, the researcher that explored in anonymization techniques said that the dataset will give highest accuracy if the dataset is being applied with the Datafly technique and SVM classifier. But as for our research, we explored that the filter and the classifier that is worked very well is Datafly technique and the J48 classifier. How this is happen? This is because the condition of our dataset and the researcher dataset is not the same. As for our dataset, it contains more on the nominal attributes in the string form such as message, loginID, firstName, lastName and so on. As for SVM classifier, it will worked poorly when the classifier is being applied on the dataset that is in the string form.

Then we can conclude that, as for our research, the optimal technique that we will used in our research is Datafly and for the base classifier, we will used J48 classifier. Datafly is the optimal technique because it will transform the data from the original data to the range of data while for J48 classifier is because it capable to handling dataset that may have errors and missing values. It also can handle both nominal and numerical attributes.

Before we conducted the experiments, we actually performed feature selection in order to determine which attributes that is important and suitable to use. So, from the selection, Info Gain Ratio Attribute and Ranker gives the optimal result where it is applied to the attributes that can take on a large number of distinct values and it often decide on the most relevant attributes. While CfsSubset Evaluation and Best-First gives the worst result where it will ignores the interaction between the instances and the classifier. So, the attributes selected will depends on the degree of redundancy of the attributes.

As for the classifier, J48 gives the optimal result while Random Tree gives the worst result. Random Tree gives the worst result because it is actually will observed to overfit for some datasets with noisy classification task. Then for the privacy technique used, Datafly gives the optimal result while AddNoise gives the worst result. AddNoise gives the worst result because it will only changes the percentage of the instance and the changes is in small amount of data. That means that not all the data will be adding the noise into it.

5.2. Discussion on Phase 1

As for Phase 2, in order to enhancing the base privacy technique, we implemented the masking technique that is proposed by Nagmode(2014)[14] to make sure the user information more secure. In the masking techniques, there are two methods that will be implemented in the dataset which are generalization and suppression methods. After the masking techniques is applied, the anonymous dataset that obtained in Phase 1 will be converted to enhancement anonymous dataset. This dataset basically more secure than anonymous dataset. And to prove the assumption before, we tested the enhancing anonymous dataset with Weka once again to see the accuracy of the dataset. Is the accuracy of enhancing anonymous dataset is more accurate compared to anonymous dataset?

The results shows that the enhancing anonymous dataset is more accurate where the comparison of the accuracy both dataset can be obtained from Figure 5 before. As we can see from the graph bar, the accuracy of the dataset with masking techniques is higher 0.10% compared to the accuracy of the dataset without applying with masking techniques. Why this happen? This is because the dataset is basically will change the actual data into the range of the data, and the data is being conceal from being known by the adversary. This shows that the enhancing anonymous dataset is more secure compared to the anonymous dataset.

From the experiment conducted again, the result shows that the accuracy of Datafly technique once again gives the optimal result which the accuracy is still maintain with 100% of accuracy.

6. Conclusion

Privacy becomes one of the main concerns in IoT due to the rapid development of technologies. If the privacy of the user is not secure enough, this will lead to the user's data that have been collected may be sold to the third party. The aim of this research is to protect the user's privacy by disabling the function of re-link the user's data to the public database.

During the Phase 1, the purposed of the experiments conducted is to define the optimal techniques that can be used in order to preserve the data mining. Based on the performance using NFC Florist dat set, the optimal privacy technique discovered is Datafly and using the same dataset, the optimal classifier that showed the most accurate result is J48 classifier.

During the Phase 2, the proposed solution in order to preserved the privacy of the user information by implementing masking techniques on the dataset that is obtained in the Phase 1. The anonymous algorithm will be implemented in the anonymous dataset and techniques that will be used in anonymous algorithm is masking techniques. It will used generalization and also suppression techniques in order to conceal the data securely without knowing the original data. The dataset obtained from Phase 2, basically is the enhancement anonymous dataset.

References

[1] Papadopoulou, E., et al., Combining pervasive computing with social networking for a student environment, in Proceedings of the Twelfth Australasian Symposium on Parallel and Distributed Computing - Volume 152. 2014, Australian Computer Society, Inc.: Auckland, New Zealand. p. 11-19.

[2] Mokhtar, S.B., L. McNamara, and L. Capra, A middleware service for pervasive social networking, in Proceedings of the International Workshop on Middleware for Pervasive Mobile and Embedded Computing. 2009, ACM: Urbana Champaign, Illinois. p. 1-6.

[3] C., W. Mobile social networking usage soars[stats]. 2010; Available from: http://mashable.com/2010/03/03/comscore-mobile-stats/.

[4] Ahn, G.-J., M. Shehab, and A. Squicciarini, Security and privacy in social networks. IEEE Internet Computing, 2011. 15(3): p. 10-12.

[5] Beach, A., M. Gartrell, and R. Han. Solutions to Security and Privacy Issues in Mobile Social Networking. in 2009 International Conference on Computational Science and Engineering. 2009.

[6] Sweeney, L., <i>k</i>-anonymity: a model for protecting privacy. Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 2002. 10(5): p. 557-570.

[7] Wen, K.S. and M.M. Singh, A Pervasive Social Networking Application: I-NFC enabled Florist Smart Advisor. IOP Conference Series: Materials Science and Engineering, 2016. 160: p. 012091.

[8] Nafa, et al., Mobile social networking applications. Commun. ACM, 2013. 56(3): p. 71-79.

[9] Verykios, V.S., et al., State-of-the-art in privacy preserving data mining. SIGMOD Rec., 2004. 33(1): p. 50-57.

[10] A., V.H.a.G., A Survey: Privacy Preservation Techniques in Data Mining International Journal of Computer Applications 2015. Volume 119 - Number 4 p. 20-26.

[11] Xie, Q. and U. Hengartner. Privacy-preserving matchmaking For mobile social networking secure against malicious users. in 2011 Ninth Annual International Conference on Privacy, Security and Trust. 2011.

[12] Ajami, R., N.A. Qirim, and N. Ramadan, Privacy Issues in Mobile Social Networks. Procedia Computer Science, 2012. 10: p. 672-679.

[13] Phalnikar, S.M.N.a.R., k-Anonymization using Multidimensional Suppression for Data De-identification. International Journal of Computer Applications, 2012. 60(11).

[14] Game, V.N.a.P., Classification Tree-Based k-Anonymity with Masking Operations to Enhance Data Utility. Proc. of Int. Conf. on Advances in Communication, Network, and Computing, CNC, 2014.

[15] Mandapati, S., R.B. Bhogapathi, and M.V.P.C.S. Rao, Classification via Clustering for Anonym zed Data. International Journal of Computer Network and Information Security, 2014. 6(3): p. 52-58.

[16] Brush, A.J.B., J. Krumm, and J. Scott, Exploring end user preferences for location obfuscation, location-based services, and the value of location, in Proceedings of the 12th ACM international conference on Ubiquitous computing. 2010, ACM: Copenhagen, Denmark. p. 95-104.

[17] Vania Bogorny, A.T.P., Paulo Martins Engel, Luis Otavio Alvares, Weka-GDPM – Integrating Classical Data Mining Toolkit to Geographic Information Systems.

[18] Luis Otavio Alvares, A.L.P., Gabriel Oliveira, Vania Bogorny, Weka-STPM: from trajectory samples to semantic trajectories.

[19] Moore Jr, R.A., CONTROLLED DATA-SWAPPING TECHNIQUES FOR MASKING PUBLIC USE MICRODATA SETS. Statistical Research Division, 1996.

[20] Manmeet Mahinderjit Singh, N.F.A.N., Privacy Preserving Techniques for Ambient Intelligence System: A Case of AMbient Intelligence Smart Home (AMISHA), in School of Computer Science. 2016, Universiti Sains Malaysia. p. 98.

[21] Madria, M.T.a.S., Sensor networks: an overview. 2003.

[22] Inan, A., M. Kantarcioglu, and E. Bertino. Using Anonymized Data for Classification. in 2009 IEEE 25th International Conference on Data Engineering. 2009.

[23] al, G.O.e., Privacy Preserving in Data Mining – Experimental Research on SMEs Data 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics 2011.G. O. Young, "Synthetic structure of industrial plastics (Book style with paper title and editor)," in *Plastics*, 2nd ed. vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15–64.

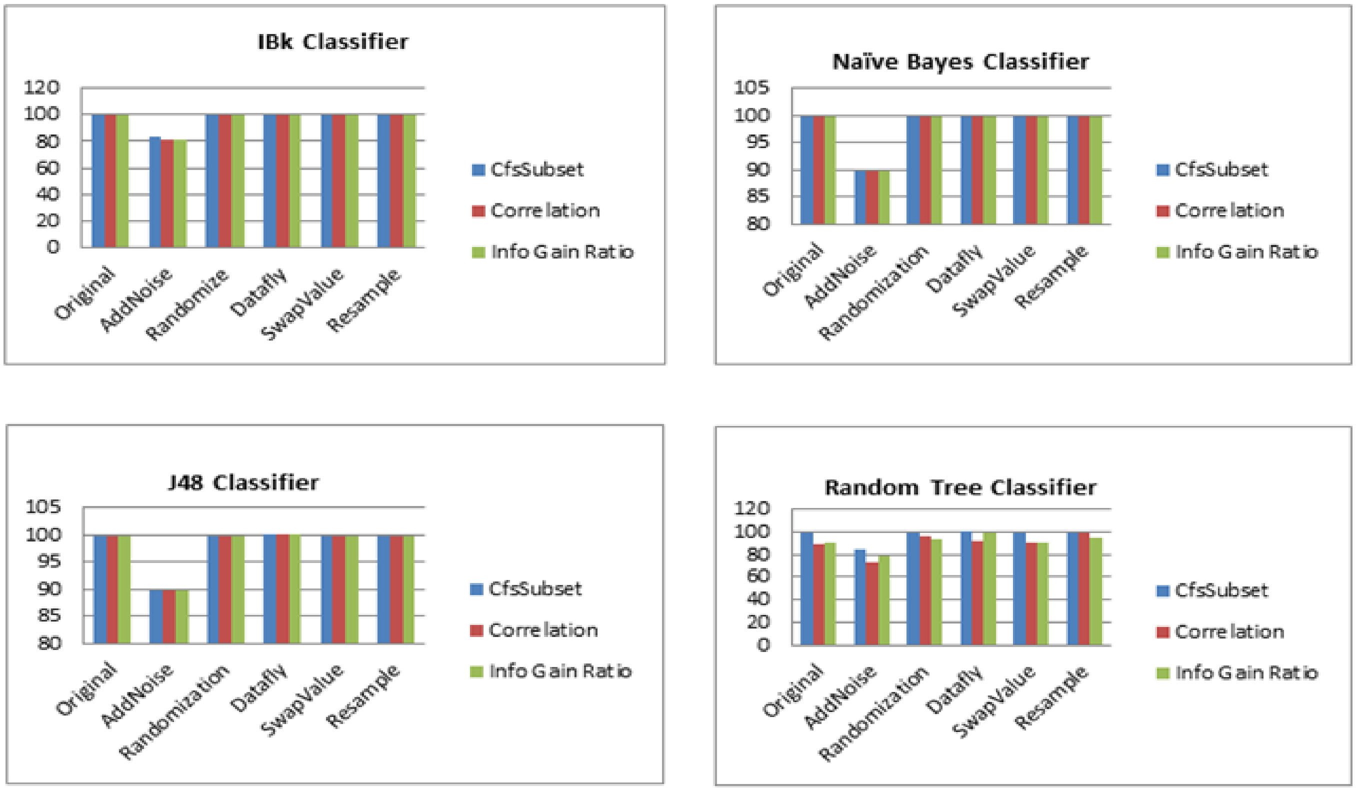
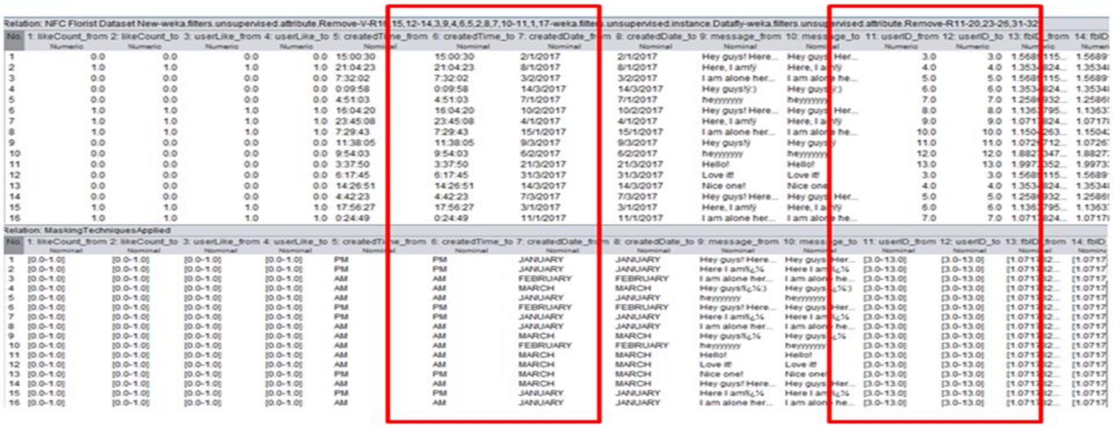


Fig. 3: Results of Accuracy on Different Classifiers with Different Feature Selections



Relation: NFC Florist Dataset New-weka.filters.unsupervised.attribute.Remove-V-R16,15,12-14,3,9,4,6,5,2,8,7,10-11,1,17-weka.filters.unsupervised.instance.Datafly-weka.filters.unsupervised.attribute.Remove-R11-20,23-26,31-32

No.	1: likeCount_from Numeric	2: likeCount_to Numeric	3: userLike_from Numeric	4: userLike_to Numeric	5: createdTime_from Nominal	6: createdTime_to Nominal	7: createdDate_from Nominal	8: createdDate_to Nominal	9: message_from Nominal	10: message_to Nominal	11: userID_from Numeric	12: userID_to Numeric	13: fbID_from Numeric	14: fbID Numeri
1	0.0	0.0	0.0	0.0	15:00:30	15:00:30	2/1/2017	2/1/2017	Hey guys! Here...	Hey guys Her...	3.0	3.0	1.568 115...	1.5689
2	1.0	1.0	1.0	1.0	21:04:23	21:04:23	8/1/2017	8/1/2017	Here, I amÿ	Here, I a ÿ	4.0	4.0	1.353 824...	1.3534
3	0.0	0.0	0.0	0.0	7:32:02	7:32:02	3/2/2017	3/2/2017	I am alone her...	I am alo he...	5.0	5.0	1.568 115...	1.5689
4	0.0	0.0	0.0	0.0	0:09:58	0:09:58	14/3/2017	14/3/2017	Hey guysÿ:)	Hey guy :)	6.0	6.0	1.353 824...	1.3534
5	0.0	0.0	0.0	0.0	4:51:03	4:51:03	7/1/2017	7/1/2017	heyyyyyy	heyyyyyy	7.0	7.0	1.258 932...	1.2586
6	1.0	1.0	1.0	1.0	16:04:20	16:04:20	10/2/2017	10/2/2017	Hey guys! Here...	Hey guys Her...	8.0	8.0	1.136 795...	1.1363
7	1.0	1.0	1.0	1.0	23:45:08	23:45:08	4/1/2017	4/1/2017	Here, I amÿ	Here, I a ÿ	9.0	9.0	1.071 824...	1.0717
8	1.0	1.0	1.0	1.0	7:29:43	7:29:43	15/1/2017	15/1/2017	I am alone her...	I am alo he...	10.0	10.0	1.150 263...	1.1504
9	0.0	0.0	0.0	0.0	11:38:05	11:38:05	9/3/2017	9/3/2017	Hey guysÿ	Hey guy ÿ	11.0	11.0	1.072 712...	1.0726
10	0.0	0.0	0.0	0.0	9:54:03	9:54:03	6/2/2017	6/2/2017	heyyyyyy	heyyyyyy	12.0	12.0	1.882 347...	1.8827
11	0.0	0.0	0.0	0.0	3:37:50	3:37:50	21/3/2017	21/3/2017	Hello!	Hello!	13.0	13.0	1.997 352...	1.9973
12	0.0	0.0	0.0	0.0	6:17:45	6:17:45	31/3/2017	31/3/2017	Love it!	Love it!	3.0	3.0	1.568 115...	1.5689
13	0.0	0.0	0.0	0.0	14:26:51	14:26:51	14/3/2017	14/3/2017	Nice one!	Nice one	4.0	4.0	1.353 824...	1.3534
14	0.0	0.0	0.0	0.0	4:42:23	4:42:23	7/3/2017	7/3/2017	Hey guys! Here...	Hey guys Her...	5.0	5.0	1.258 932...	1.2586
15	1.0	1.0	1.0	1.0	17:56:27	17:56:27	3/1/2017	3/1/2017	Here, I amÿ	Here, I a ÿ	6.0	6.0	1.136 795...	1.1363
16	1.0	1.0	1.0	1.0	0:24:49	0:24:49	11/1/2017	11/1/2017	I am alone her...	I am alo he...	7.0	7.0	1.071 824...	1.0717

Relation: MaskingTechniquesApplied

No.	1: likeCount_from Nominal	2: likeCount_to Nominal	3: userLike_from Nominal	4: userLike_to Nominal	5: createdTime_from Nominal	6: createdTime_to Nominal	7: createdDate_from Nominal	8: createdDate_to Nominal	9: message_from Nominal	10: message_to Nominal	11: userID_from Nominal	12: userID_to Nominal	13: fbID_from Nominal	14: fbID Nomina
1	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	PM	PM	JANUARY	JANUARY	Hey guys! Here...	Hey guys Her...	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
2	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	PM	PM	JANUARY	JANUARY	Here I amï¿½	Here I ar ¿½	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
3	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	FEBRUARY	FEBRUARY	I am alone her...	I am alon he...	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
4	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	MARCH	MARCH	Hey guysï¿½:)	Hey guys ¿½:)	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
5	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	JANUARY	JANUARY	heyyyyyyy	heyyyyyyy	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
6	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	PM	PM	FEBRUARY	FEBRUARY	Hey guys! Here...	Hey guys Her...	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
7	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	PM	PM	JANUARY	JANUARY	Here I amï¿½	Here I ar ¿½	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
8	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	JANUARY	JANUARY	I am alone her...	I am alon he...	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
9	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	MARCH	MARCH	Hey guysï¿½	Hey guys ¿½	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
10	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	FEBRUARY	FEBRUARY	heyyyyyyy	heyyyyyyy	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
11	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	MARCH	MARCH	Hello!	Hello!	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
12	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	MARCH	MARCH	Love it!	Love it!	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
13	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	PM	PM	MARCH	MARCH	Nice one!	Nice one!	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
14	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	MARCH	MARCH	Hey guys! Here...	Hey guys Her...	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
15	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	PM	PM	JANUARY	JANUARY	Here I amï¿½	Here I ar ¿½	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717
16	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	[0.0-1.0]	AM	AM	JANUARY	JANUARY	I am alone her...	I am alon he...	[3.0-13.0]	[3.0-13.0]	[1.071 82...	[1.0717

Fig 4 : Data Collected Before and After Masking Techniques Applied

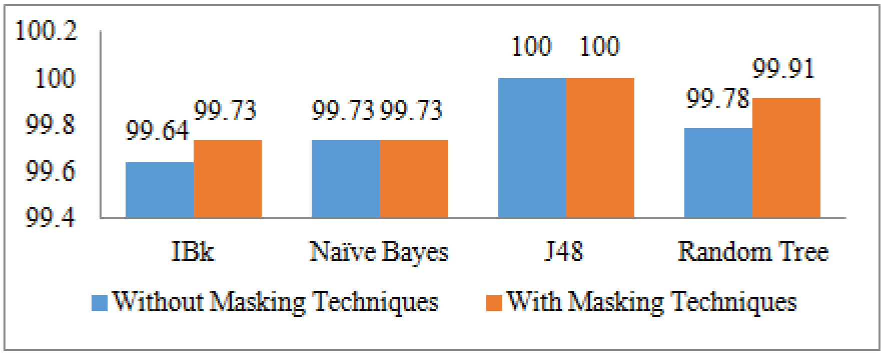


Fig. 5: Accuracy of Datafly Technique using Different Classifiers Before and After Masking Techniques Applied

Table 5 : Summary and Weighted Average Before and After Masking Technique Applied

Measurements	Before Masking Technique Applied				After Masking Technique Applied			
	Datafly+Naïve Bayes	Datafly+IBk	Datafly+J48	Datafly+Random Tree	Datafly+Naïve Bayes	Datafly+IBk	Datafly+J48	Datafly+Random Tree
KAPPA	0.994	0.992	1.000	0.866	1.000	0.997	1.000	0.998
MAE	0.002	0.004	0.000	0.069	0.001	0.002	0.000	0.001
RMSE	0.051	0.059	0.000	0.241	0.002	0.039	0.000	0.032
RAE (%)	0.541	0.821	0.000	13.991	0.244	0.308	0.000	0.201
RRSE (%)	10.406	12.010	0.000	48.556	0.283	7.754	0.000	6.332
TPR	0.997	0.996	1.000	0.934	1.000	0.999	1.000	0.999
FPR	0.003	0.004	0.000	0.067	0.000	0.002	0.000	0.001
P	0.997	0.996	1.000	0.934	1.000	0.999	1.000	0.999
R	0.997	0.996	1.000	0.934	1.000	0.999	1.000	0.999
FM	0.997	0.996	1.000	0.934	1.000	0.999	1.000	0.999
ROC	**1.000**	**0.997**	**1.000**	**0.949**	**1.000**	**1.000**	**1.000**	**0.999**