Human-centric Computing and Information Sciences

RESEARCH Open Access

An information-aware visualization for privacy-preserving accelerometer data sharing

Fengjun Xiao[1,2], Mingming Lu[3*], Ying Zhao[3*], Soumia Menasria[3], Dan Meng[3], Shangsheng Xie[3], Juncai Li[3] and Chengzhi Li[1,2]

*Correspondence: mingminglu@csu.edu.cn; zhaoying@csu.edu.cn
[3] School of Information Science and Engineering, Central South University, Changsha, China
Full list of author information is available at the end of the article

Abstract

In the age of big data, plenty of valuable sensing data have been shared to enhance scientific innovation. However, this may cause unexpected privacy leakage. Although numerous privacy preservation techniques, such as perturbation, encryption, and anonymization, have been proposed to conceal sensitive information, it is usually at the cost of the application utility. Moreover, most of the existing works did not distinguished the underlying factors, such as data features and sampling rate, which contribute differently to utility and privacy information implied in the shared data. To well balance the application utility and privacy leakage for data sharing, we utilize mutual information and visualization techniques to analyze the impact of the underlying factors on utility and privacy, respectively, and design an interactive visualization tool to help users identify the appropriate solution to achieve the objectives of high application utility and low privacy leakage simultaneously. To illustrate the effectiveness of the proposed scheme and tool, accelerometer data collected from mobile devices have been adopted as an illustrative example. Experimental study has shown that feature selection and sampling frequency play dominant roles in reducing privacy leakage with much less reduction on utility, and the proposed visualization tool can effectively recommend the appropriate combination of features and sampling rates that can help users make decision on the trade-off between utility and privacy.

Keywords: Accelerometer, Activity recognition, Application utility, Data sharing, Mutual information, Privacy preserving, Visualization

Introduction

As MEMS sensors are ubiquitously embedded into mobile devices, such as smartphones, wearable devices, and tablets, a large amount of sensing data have been generated daily through these embedded sensors. Most recently, some pioneer institutions [1–4] have begun to share a large amount of sensing data collected from thousands of mobile devices, which have given a strong impetus to the development of numerous novel applications, such as smart cities [5], healthcare [6], and activity recognition [7]. Although sensing data sharing have enhanced scientific innovation, it may incur unexpected privacy leakage.

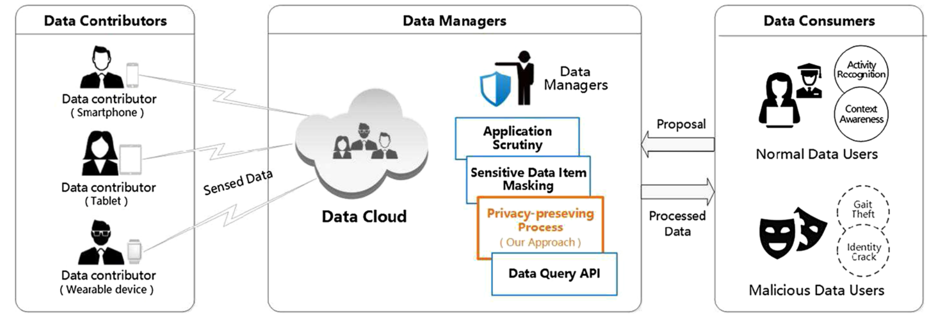


Fig. 1 A typical scenario of sensing data sharing

To illustrate the privacy leakage problem associated with accelerometer data sharing, we illustrate a typical scenario of accelerometer data sharing as shown in Fig. 1. In this scenario, data contributors send the data collected from their mobile devices (such as smartphones, tablets, and wearable devices) to credible third parties (such as government agencies and research communities), who are responsible for the data management. Once the data have been collected to the data managers, data consumers can submit their proposals to the data managers for the data analysis requests with the claimed usage. After a fully scrutiny on the proposals, data managers would process raw data through various privacy-preserving techniques. Finally, the data query API will be provided to the approved data consumers for data analysis.

However, the existing practical privacy-preserving approaches, such as data perturbation [8], encryption [9], and anonymization [10] may not be able to fully filter out the possibility of privacy leakage, because malicious data users may infer the private information of data contributors, even though those information have been perturbed, encrypted, or anonymized. The fundamental reason is that the shared data may statistically contain privacy information that can be inferred through various machine learning models.

To combat the statistical inference attacks, two information-based privacy paradigms (differential privacy [11] and inferential privacy [12, 13]) have been proposed. Differential privacy intends to protect presence privacy through mathematical deduction, i.e., the presence or absence of an individual data record cannot be distinguished with arbitrary accuracy. However, if there exists correlation among data records, the attributes of absent data records can still be inferred. To avoid this issue, inferential privacy has been proposed to avoid any private attribute to be statistically inferred.

Although inferential privacy can theoretically guarantee no privacy leakage, it cannot directly apply to the application scenario of accelerometer data sharing due to the following reasons: (1) it does not consider the utility-privacy trade-off, as a consequence of which, the claimed privacy target may compromise the intended application utility (a concrete scenario can be referred to "Extension and discusssion"); (2) it does not take into account users' preferences, which have been illustrated to be subjective, i.e., different users may have different privacy concerns [14]; (3) it is a privacy protection principle, which does not provide concrete methods to analyze the impacting factors that can affect the utility-privacy trade-off.

To address the above three issues, we first consider an explanatory scenarios, where the application usage claimed by the data consumers is activity recognition and the potential privacy leakage is user identity, because activity recognition [15] is the building block for a lot of applications, such as sporting, sleeping, and health, and identity leakage is a serious privacy leakage problem with significant impact. Then, we examine the key factors that may impact the application utility (the accuracy of activity recognition) and the privacy leakage (the accuracy of identity recognition) associated with accelerometer data. Based on extensive experiments conducted through various machine learning models, we identified that data features and sampling rates play dominant roles in determine the accuracy for both activity recognition and identity recognition.

According to the principle of the inferential privacy, to avoid the dependence on a particular learning model, we proposed a mutual information based feature selection method and the associated sampling rate adjustment scheme to quantify the utility and privacy information implied by each feature and the corresponding sampling rates associated with each feature.

To include users' preferences into privacy-preserving data sharing, an interactive visualization tool is designed to enable users to observe and analyze the utility and privacy information implied by the combinations of features and sampling rates. Several interactive visualization modules and the recommended solution module help users to identify the appropriate combination of features and sampling rates that may satisfy their preferred high-utility and low-privacy goal. Intensive experiment evaluations on more than one datasets have shown that the proposed visualization scheme can effectively protect users' privacy information other than the identity privacy and can still achieve a high application utility. Moreover, the mutual information between various data attributes (such as gender, height, weight, and etc) and each underlying factor (including all features and sampling rates) are extensively evaluated through various experiments, from which it can be concluded that the proposed visualization based privacy-preserving scheme can extend to scenarios with the application utility and privacy other than activity recognition and identity recognition. Furthermore, a valuable insight on the design space is discussed for future study.

The contributions of this work can be enumerated as follows: (1) we identified the underlying key factors contributing to the utility and privacy, respectively; (2) we adopted a mutual information based scheme to identify the privacy-preserving data sharing solution that can achieve both high utility and low privacy simultaneously; (3) we proposed an interactive visualization tool that takes users' preference into account to obtain a customized privacy-preserving data sharing solution; (4) we shed a light on a potential large design space for information-aware privacy preserving data sharing.

The rest of this work is organized as follows. "Related works" presents the works most related to our work. "Factor analysis for the utility and privacy of accelerometerdata sharing" analyzes the underlying factors contributing to utility and privacy information, respectively, through experimental study. In "Mutual information based feature selection and sampling rateadjustment", a mutual information based feature selection and sampling rate adjustment scheme have been proposed. "Visualization based privacy-preserving scheme" describes the interactive visualization tool designed to provide a customized privacy preserving data sharing solution. "Evaluation through case study" describes the process to use VISEE through an example, verifies the proposed work through another data set, extends the application scenarios, and discusses about the design space of the proposed privacy data sharing scheme. Finally, "Conclusion" concludes this work and describes the future work.

Related works

To protect the privacy of the shared data, numerous privacy-preserving works had been proposed. Most of the existing works applied perturbation, encryption, and anonymization to the raw data.

Data perturbation is the most common and direct privacy protection method, which hides the privacy through applying random noise on the raw data. However, it may leak privacy information through statistical inference, even though the perturbed data is hard to be reconstructed [8, 16–23]. However, many limitations exist for perturbation. Perturbed data still maintain a lot of characteristics of the raw data, from which the malicious users may infer privacy information. Additionally, it requires different random algorithms to perturb different types of data. As a result, this introduces additional cost to estimate the level of privacy protection for each random algorithm. Finally, it incurs additional preprocess overheads for data users.

Data encryption implements privacy protection through applying encryption techniques to the raw data, which can still supports certain computations for learning [24–30]. Although data encryption has a strong capability to protect privacy, it is at the cost of high computation complexity. Additionally, it is unable to quantify the amount of private information to be protected. Therefore, it is hard to evaluate the performance of the encryption-based schemes in terms of the amount of protected privacy information.

Data anonymization achieved privacy protection through selecting the data to be shared or hashing sensitive information. k-anonymity partitions data into public properties and sensitive properties and requires that at least k undistinguished data exist for each equivalent class in the shared data set, so that malicious users cannot identify any individual data contributor based on the sensitive information from the other $k - 1$ pieces of data [31]. However, malicious users can infer any individual's privacy if any equivalent class contains sensitive attributes because k-anonymity does not consider the distribution of the sensitive properties. l-diversity solves this problem by requiring that at least l different sensitive attributes exist for each equivalent class [32]. However, l-diversity ignores the globalness of the data distribution, which can be utilized by malicious users if certain sensitive attributes occur frequently in an equivalent class. To address this, t-closeness [10] has been proposed to enforce the distribution of sensitive attributes within each equivalent class to be close to the global distribution as much as possible.

However, the above privacy protection schemes cannot resist the inference attack, because they did not quantify the privacy leakage level through information theory. Moreover, these schemes assumed that malicious users did not have background knowledge about the data. To address these issues, Dwork et al. proposed differential privacy [11], which utilized mathematical models to guarantee the statistical properties that the presence or absence of a user's data cannot be statistically distinguishable with arbitrary accuracy, even though malicious users knew the background information of

Table 1 The comparison between the related works and our work

Models	FactorInvest	InfoTheory	VisualAnalysis	Utility-privacy
Perturbation	No	No	No	No
Encryption	No	No	No	No
Anonymization	No	No	No	No
DifferPrivacy	No	Yes	No	No
InferPrivacy	No	Yes	No	No
FactorInvest	Yes	No	No	No
FeatureSelect	No	No	No	No
PreviousWork	Yes	No	No	Yes
ThisWork	Yes	Yes	Yes	Yes

the shared data. However, differential privacy cannot withhold the inference attack if a user's data is correlated with other users' data. Although inferential privacy [12, 13] can address the data correlation issue, it cannot directly handle the trade-off between utility and privacy, since it focused on protecting the privacy without considering the application utility. Moreover, inferential privacy did not take the end users' preference into account and did not consider the utility/privacy information implied by each underlying factor.

Some recent works [33, 34] have also investigated the factors affecting users choices toward the disclosure of their personal data. Their works are different from our work, which considered the factors from the perspectives of data features and sampling rates. Although the existing works [35, 36] have utilized mutual information for feature selection, the major difference of our work from theirs lies in that our work utilizes mutual information to identify the set of features and sampling rates that balance the trade-off between the utility maximization and the privacy minimization for the sensing data sharing scenario, instead of computing the set of features that simply maximizing model accuracy. Our previous work [37] has already considered the scenarios of accelerometer data sharing. However, this previous work did not conduct extensive experiments to evaluate the effects of features and sampling rates, propose mutual information based feature selection and sampling rate adjustment scheme, or design an interactive visualization tool for customized privacy-preserving data sharing.

To illustrate the difference between previous works and this work, we enumerate their differences in Table 1, where DifferPrivacy, InferPrivacy, FactorInvest, InfoTheory, and Utility-privacy represent the differential privacy model, the inferential privacy model, the factor investigation models/methods, the information theory based techniques, and the utility-privacy trade-off, respectively.

Factor analysis for the utility and privacy of accelerometer data sharing

To identify whether the data manager can provide the accelerometer data to the data consumers for the activity recognition application without leaking the identity information, it is necessary to carefully examine the accelerometer data. An accelerometer simply senses the accelerations of three orthogonal direction as shown in Fig. 2. The underlying reason that accelerometer data can be used to infer a mobile device user's activity is that different activities, such as stay, walk, and jog, show different characteristic in the sensed

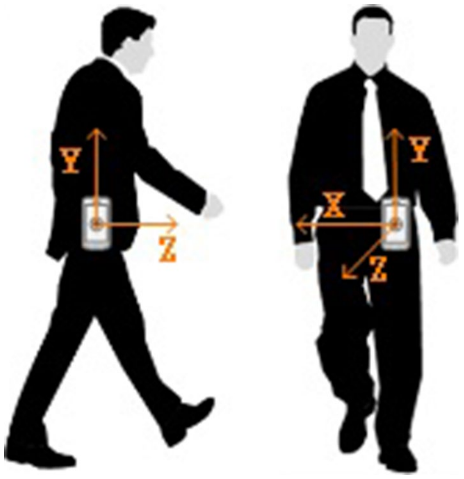


Fig. 2 The sensing directions of the 3-dimensionalacclerometer accelerometer relative to a walking person [1]

accelerometer data. Similarly, a user's identity can be inferred due to the gait characteristic depending on a user's muscle growth, bone structure, height, weight, and etc.

Previous studies on activity recognition [15, 38] and identity recognition [39, 40] have identified three key factors, namely learning model, feature selection and data sampling rate, which have significant influence on the recognition accuracy [41]. Because data managers have no idea about the models to be used by data consumers, the remaining two factors should be focused to analyze the possibility of privacy-preserving data sharing.

In the following, two sets of experiments will be conducted to evaluate the effects of the feature selection and sampling rate adjustment, respectively, on the accuracy of both activity and identity recognitions. In both of the experiment sets, two acceleration datasets were selected from the HASC corpus dataset [1]: one data set for activity recognition, consisted of the acceleration data from 100 volunteers with six activity types, namely stay, walk, jog, skip, stair-up, and stair-down, and the other one for identity recognition, which contains the acceleration data from 100 subjects with activity type being walk, as walking is known as the most common and stable human activity for distinguishing individuals [42].

The effect of feature selection

Through carefully examination of the existing features used in the related works, 12 typical features were chosen for our experiments. These features can be divided into two categories: time domain and frequency domain. Each feature was extracted from three acceleration directions respectively. As a result, there are 36 features in total, as shown in Table 2. Before feature extraction, we partitioned the time series of the acceleration data into multiple 1-s segments, since the average gait cycle time is about 1-s for people at the age from 18 to 49 [15]. From each segment, the 36 features were extracted to obtain a feature vector with activity and identity labels.

Table 2 Feature list

Category	Feature names for 3 directions	Feature description
Time	Xaverage,Yaverage, Zaverage	The average of each segment
	Xvariance, Yvariance, Zvariance	The variance of each segment
	Xstandarddev,Ystandarddev, Zstandarddev	The standard deviation of each segment
	Xaverageabsdif,Yaverageabsdif, Zaverageabsdif	The average of the absolute deviations of each segment
	Xnanmax, Ynanmax, Znanmax	The maximum values of each segment
	Xnanmin, Ynanmin, Znanmin	The minimum values of each segment
	Xmedian, Ymedian, Zmedian	The 50th percentile values of each segment
	Xperseven, Yperseven, Zperseven	The 70th percentile values of each segment
	Xpereight, Ypereight, Zpereight	The 80th percentile values of each segment
	Xpernine, Ypernine, Zpernine	The 90th percentile values of each segment
Frequency	Xpeakamplitude, Ypeakamplitude, Zpeakamplitude, Xdomfre,Ydomfre, Zdomfre	The value of the waveform's peak of each segment after Fourier Transform
	Xenergy, Yenergy, Znergy	The power of the signal of each segment after Fourier Transform

Table 3 The accuracy of activity recognition on the HASC data set through various machine learning models

Model	10 (%)	30 (%)	50 (%)	Average accuracy (%)
RF	95.1348	94.4644	93.7216	94.4403
J48	89.8845	89.2323	88.2606	89.1258
NB	70.0735	70.136	69.4085	69.8727
BN	84.1267	80.5552	78.1867	80.9562
LMT	91.1271	91.2667	90.4179	90.9372
MP	92.6671	90.1427	88.3929	90.4009
IB3	94.5747	93.0932	92.5717	93.4616
Logistic	87.9769	85.5513	83.8274	85.7852
RN	78.4389	77.7565	75.7513	77.3157
DT	78.3514	79.583	78.0883	78.6742

Underlined values indicate the best accuracy result achieved by a certain model among all available models

In the first set of experiments, 10 feature combinations were prepared for the evaluation, where 5 feature combinations were manually selected by domain experts and the other 5 feature combinations were recommended by Weka [38] according to different selection strategies. For each feature combination, the results for both activity and identity recognitions were obtained by 10 classification models in Weka with default parameters and tenfold cross validation. The 10 classification models are NB (Nave Bayes), J48 (one method of decision trees), IB3 (instance-based learning), BN (Bayes Net), Logistic, LMT (Logistic model trees), RN (RBFNetwork), DT (Decision Table), MP (Multilayer Perceptron), and RF (Random Forest), the detailed descriptions of which can be found in Bishop's famous book on machine learning [43].

To compare those models, 10, 30, 50 users were randomly chosen to compute the accuracy for both activity recognition and identity recognition. The experiment results for activity recognition and identity recognition are shown in Tables 3 and 4, respectively, from which it can be observed that the RF model has the highest accuracy for

Table 4 The accuracy of identity recognition on the HASC data set through various machine learning models

Model	10 (%)	30 (%)	50 (%)	Average accuracy (%)
RF	98.0851	95.8194	93.7983	95.9009
J48	95.2128	85.8194	79.9105	86.9809
NB	97.6596	91.7057	90.6669	93.3441
BN	97.8723	94.6823	91.2973	94.6173
LMT	98.5106	96.8562	95.2013	96.8560
MP	98.5106	96.087	95.181	96.5929
IB3	97.5532	95.4515	93.3103	95.4383
Logistic	98.7234	94.2809	91.5819	94.8621
RN	97.6596	94.6823	92.1322	94.8247
DT	80.5319	60.5351	49.7967	63.6212

Underlined values indicate the best accuracy result achieved by a certain model among all available models

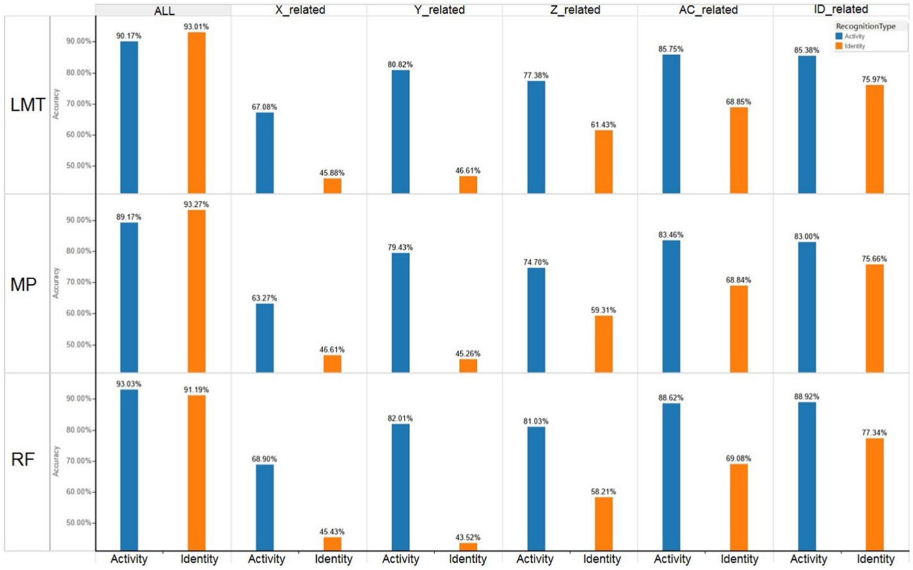


Fig. 3 The accuracy comparison for different sets of features through various models

activity recognition, while LMT and MP are the top 2 models with the highest accuracy for identity recognition. Therefore, in the following, models RF, LMT, and MP are chosen to compute accuracy for both activity and identity recognitions.

The experiment results for the LMT, MP and RF models with 5 feature combinations, namely, All (all 36 features), X_related (12 features associated with X-axes), Y_related (12 features associated Y-axes), Z_related (12 features associated with Z-axes), AC_related (the combination of the 12 feature recommended for activity recognition by the CfsSubsetEva attribute evaluator and the BestFirst search method in Weka), ID_related (the combination of the 12 feature recommended for identity recognition by the InfoGainAttributeEval attribute evaluator and the Ranker search method in Weka), are shown in Fig. 3.

From Fig. 3, four observations can be directly obtained. First, 'All' achieves the highest accuracies on both activity and identity recognition. Second, Y_related is better than

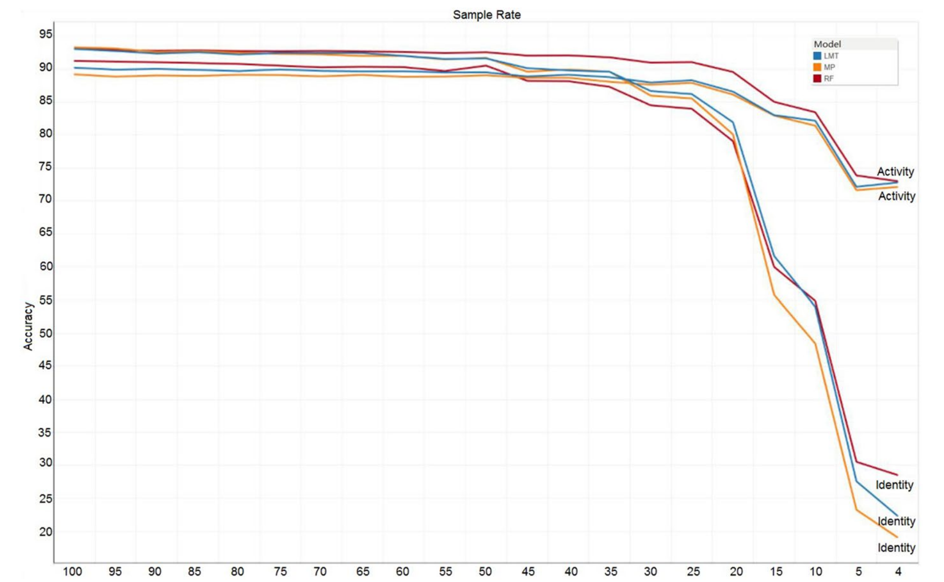


Fig. 4 The impact of the sampling rate on the accuracies of both activity and identity recognitions

X_related and Z_related on activity recognition in terms of accuracy, while Z_related is a much better choice than X_related and Y_related for identity recognition. Third, even though ID_related are selected for identity recognition, it surprisingly performs well for activity recognition. Fourth, it is possible to identify a combination of features, such as Y_related, which can achieve high accuracy for activity recognition and low accuracy for identity recognition simultaneously.

In a summary, the above experiment results demonstrate that the feature selection can dramatically influence the accuracy of both activity and identity recognitions. This inspires us to appropriately choose a feature set, which can achieve a good balance between utility (in term of the activity recognition accuracy) and privacy (in term of the identity recognition privacy) by maximizing the utility and minimizing the privacy simultaneously.

The effect of sampling rate

The second set of experiments is used to observe the effect of the sampling rates on both activity recognition and identity recognition in term of accuracy. The sampling rate of the experiment data is 100 Hz. In our experiments, the sampling rate is set as a hyper-parameter, which is gradually reduced from 100 to 4 Hz. For each value of the sampling rate, 36 features are extracted to feed into the three classification models, LMT, MP, and RF for performance comparison. Figure 4 shows the experiment results, from which it can be observed that the accuracies for both activity and identity recognition do not change much when the sampling rate reduced from 100 to 35 Hz, and the accuracy of identity recognition declines much faster than activity recognition when the sampling rate reduces from 35 to 4 Hz.

Overall, the experiment results tell us that the deduction of sampling rate can reduce the accuracy for both activity and identity recognitions, and moreover, the accuracy reduction for the identity recognition will be much faster than activity recognition.

Thus, sampling rate is a crucial ingredient for designing a privacy-preserving data sharing scheme.

From the two sets of experiments, we can conclude that both feature selection and sampling rate adjustment can be used to balance the trade-off between utility (activity recognition) and privacy (identity recognition). It is straightforward to come up with the idea to combine the feature selection and sampling rate adjustment to achieve a better result for privacy-preserving data sharing.

Mutual information based feature selection and sampling rate adjustment

Experiments shown in "Factor analysis for the utility and privacy of accelerometerdata sharing" have demonstrated that the combination of the feature selection and the sampling rate adjustment may enable a good privacy-preserving accelerometer data sharing scheme. However, in the real scenarios of data sharing, it is impossible to know in advance about the learning models used by the data consumers. Therefore, it is not appropriate to simply enumerate all the existing learning models to evaluate the appropriate combinations of feature selection and sampling rate adjustment. Instead, it is desirable to define a metric to quantify the amount of information contained in the shared data, so that the shared data can achieve a good balance between the utility and the privacy information through an appropriate combination of the feature selection and sampling rate adjustment.

Previous studies [44, 45] have shown that the classification type information implied by a feature can be used to evaluate the usefulness of a feature. Therefore, the mutual information between features and classification types is a good metric to measure the utility/privacy information implied by data, since mutual information can be used to quantify the information correlation between the input and output of a classification model. More specifically, a classification model for activity recognition can be used to measure the utility information in the input data, while a model for identity recognition can measure the privacy information in the data. Moreover, as mutual information originates from Shannon's information theory, which is independent of the specifical learning models, the utility/privacy information measured through mutual information can provide an upper/lower bound in term of accuracy for all learning models, including models not yet invented.

In the following, we will first describe the computation of the mutual information between different sets of features and the two classification models (i.e., activity recognition and identity recognition). Then, an algorithm will be proposed to select the set of features with high contributions to activity recognition but low contributions to identity recognition. After that, we will describe the computation of mutual information between different sampling rates and the two classification models to choose the appropriate sampling rate with high utility information but low privacy information.

To facilitate the mutual information calculation, we formalize the definition of mutual information as follows. Let F be a discrete random variable with n states, $f_i(i = 1 \ldots n)$, where f_i denotes a data feature, and C be a discrete random variable with m value, $c_j(j = 1 \ldots m)$, where c_j represents a classification type. The mutual information between the feature variable F and the category variable C is as follow.

$$I(F, C) = H(C) - H(C|F), \tag{1}$$

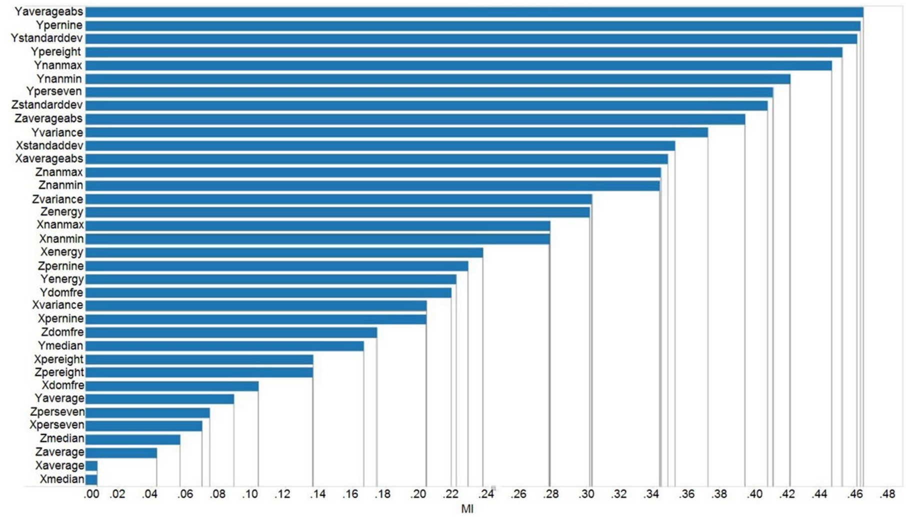


Fig. 5 The ordered mutual information between each individual feature and activity types

where $H(C)$ is the entropy of the classification type set, and $H(C|F)$ is the conditional entropy of the classification type set if the feature set F is known.

Mutual information based feature selection

Let $C_{activity}$ denote the activity type. The entropy $H(C_{activity})$ is used to measure the uncertainty about the activity types, while $H(C_{activity}|F)$ evaluates the uncertainty about the activity types if the set of features is known. From Eq. (1), it can be inferred that mutual information $I(F, C_{activity})$ actually measures the reduced uncertainty of the activity types if the feature set F has been provided. Therefore, mutual information can be used to measure the information concerning the activity types implied by the feature set. In other words, mutual information is sufficient to evaluate the capability of a feature set to predict activity types.

Usually, it is not convenient to directly apply Eq. (1) to calculate the mutual information. Instead, the following equation is used to calculate the mutual information.

$$I(F, C_{activity}) = \sum_{i,j} p(f_i, c_j) \log \frac{p(f_i, c_j)}{p(f_i)p(c_j)}, \tag{2}$$

where $p(f_i)$ and $p(c_j)$ are the probability of feature f_i and activity type c_j, respectively, while $p(f_i, c_j)$ is the joint probability of f_i and c_j.

Based on the above discussion, to evaluate the contribution of each individual feature f_i to predict activity types, we can simply calculate $I(f_i, C_{activity})$ based on Eq. (2). In our experiments, 100 users' accelerometer data from the HASC data set [1] with 6 activity types (namely stay, walk, jog, skip, stair-up, and stair-down) under the sampling rate 100 Hz are used to calculate the mutual information associated with each individual feature. Figure 5 enumerates the ordered mutual information between each feature and activity types, where the horizontal axis presents the mutual information and the

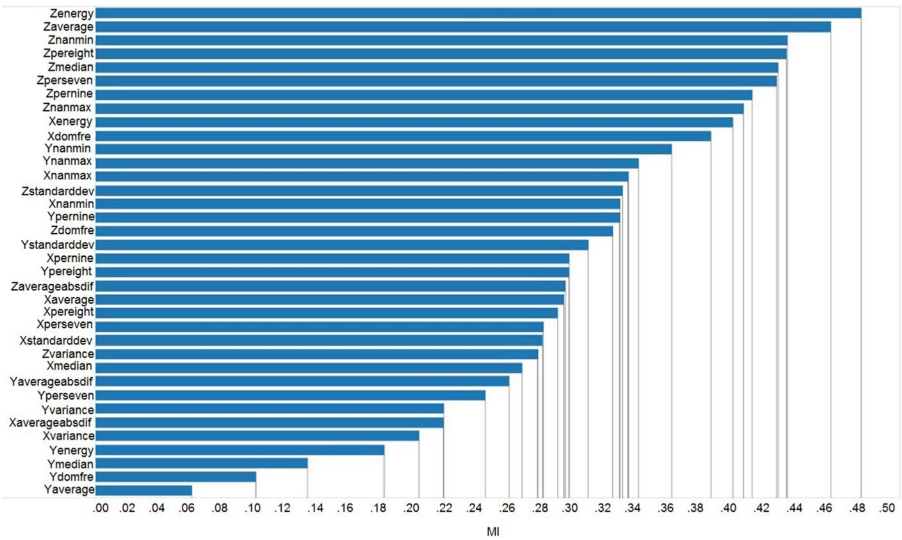


Fig. 6 The ordered mutual information between each individual feature and identities

vertical axis denotes the feature names. From Fig. 5, it can be observed that the top 7 features are associated with the Y axis, NO. 8 and NO. 9 are features related to Z axis, NO. 10 is related to Y axis again, and X axis related features have lower ranks. These results are consistent with the experiment results shown in "The Effect of feature selection", i.e., Y_related features is better than X_related and Z_related features in terms of activity type prediction accuracy.

Similarly, the mutual information between the features and the identities can be calculated as follows.

$$I(F, C_{identity}) = \sum_{i,j} p(f_i, c_j) \log \frac{p(f_i, c_j)}{p(f_i)p(c_j)}, \tag{3}$$

Also, it is easy to evaluate the contribution of each individual feature f_i to predict identities through the calculation of $I(f_i, C_{identity})$ based on Eq. (3). 100 users' walking accelerometer data from the HASC data set under the sampling rate 100 Hz are used to calculate the contribution of each individual feature. Figure 6 enumerates the ordered mutual information between each feature and identities, where the horizontal axis presents the mutual information and the vertical axis denotes the feature names. From Fig. 6, it can be observed that the top 8 features are related to the Z axis, NO. 9 and NO. 10 features are related to X axis, and Y axis related features have lower ranks instead. These results are consistent with the experiment results in "The Effect of feature selection", i.e., Z_related features is more important to identity recognition.

However, the purpose of this work is to identify the set of features with high activity recognition accuracy and low identity recognition accuracy. From the utility and privacy information ranking shown in Figs. 5 and 6, respectively, two feature sets with the top mutual information associated with activity types and identities can be choses as follows: $F_1 = \{$Yaverageabsdif, Ypernine, Ystandarddev, Ypereight, Ynanmax, Ynanmin,

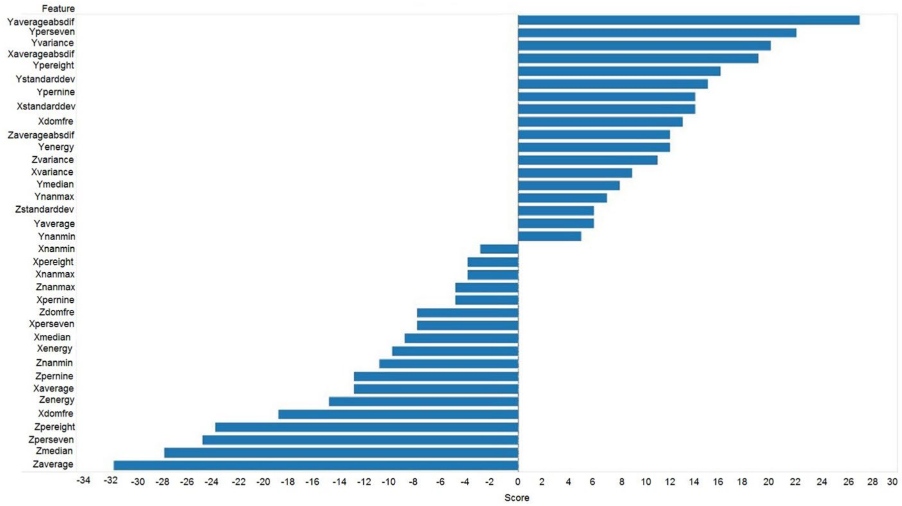


Fig. 7 MFS based feature ordering

Yperseven, Zstandarddev, Zaverageabsdif, Yvariance, Xstandarddev, Xaverageabsdif}, and $F_2 = \{$Zenergy, Zaverage, Znanmin, Zpereight, Zmedian, Zperseven, Zpernine, Znanmax, Xenergy, Xdomfre, Ynanmin, Ynanmax$\}$. Through the comparison of the two feature sets F_1 and F_2, it can be observed that $F_1 \cap F_2 = \{Ynanmin, Ynanmax\}$. These two features actually play a role in increasing the recognition accuracy for both activity and identity. Thus, these two features should be avoided for the purpose of privacy-preserving data sharing. To select the features satisfying this purpose, we propose a Mutual-information based Feature Selection (MFS) algorithm.

The key idea of MFS is as follows: for any feature f_i, if $I(f_i, C_{activity})$ is large, then f_i will be assigned a high accepting weight $Accept(f_i)$; if $I(f_i, C_{identity})$ is large, then f_i will be assigned a high rejecting weight $Reject(f_i)$. Whether a feature meets the high utility and low privacy requirement depends on its corresponding score, calculated as follows.

$$Score(f_i) = \alpha \times Accept(f_i) - \beta \times Reject(f_i), \tag{4}$$

where α and β are weight parameters. Finally, all features will be ordered according to their score rankings, based on which the features with high rankings will be selected.

Based on the MFS algorithm, $Score(f_i)$ for each f_i is computed and all the features are ordered as shown in Fig. 7. Based on this feature ranking, MFS selects a feature set $F_3 = \{$Yaverageabsdif, Yperseven, Yvariance, Xaverageabsdif, Ypereight, Ystandarddev, Ypernine$\}$.

Mutual information based sampling rate adjustment

Similar to feature selection, sampling rate adjustment can also be evaluated through mutual information. More specifically, when a feature can contribute significantly to activity/identity recognition, accelerometer data sampled with finer granularity may reflect more characteristic of the corresponding activity/identity types. Otherwise, the feature with less sampling rate may contain much less utility/privacy information. Therefore, in the following, the mutual information between each feature and the activity/

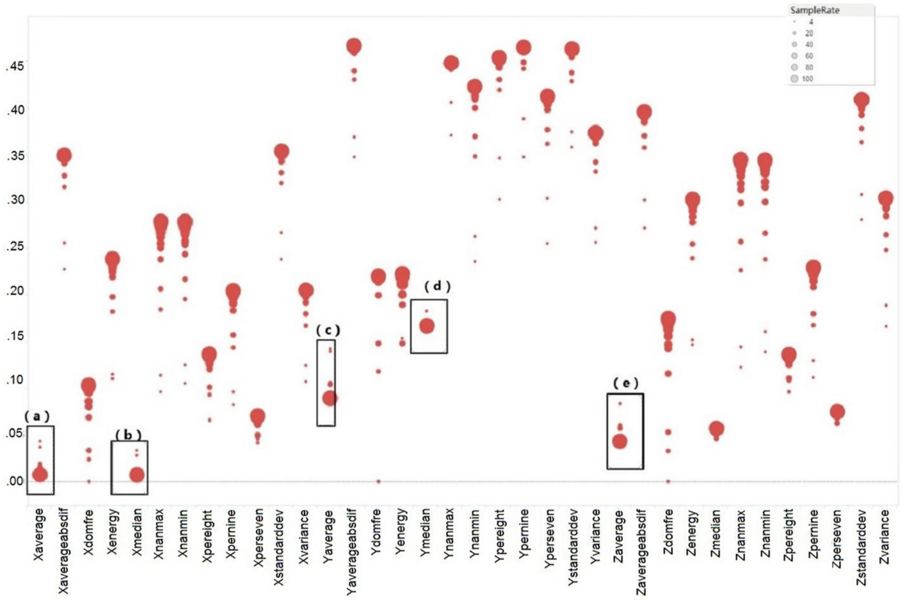


Fig. 8 The variation of utility information along with the change of the sampling rate

identity type set under different sampling rates will be calculated to observe the variation trend of the mutual information along with the reducing sampling rate.

To observe the variation of mutual information $I(f_i, C_{activity})$ along with the adjustment of the sampling rate, 100 users' accelerometer data labelled with the 6 activity types are used to calculate the mutual information for each sampling rate gradually reduced from 100 to 4 Hz. Figure 8 illustrates the mutual information variation for each feature along with the change of sampling rate. In Fig. 8, the horizontal axis represents features, the vertical axis denotes the mutual information, and a red circle means the mutual information for a given pair of feature and activity type set, with the size of the circle denoting the sampling rate (the larger the circle, the larger the sampling rate).

From Fig. 8, it can be observed that the mutual information for most features, except Xaverage, Xmedian, Yaverage, Ymedian, and Zaverage [highlighted with black boxes label with (a)–(e), respectively], decrease along with the reducing sampling rate. Actually, these 5 exceptional features provide a relatively low information to activity recognition. It can also be inferred from Fig. 8 that the features with higher rankings at the high sampling rate usually also have higher rankings under the low sampling rate.

The mutual information between the features and the identities also shows a significant trend, as illustrated through the experiment results shown in Fig. 9, from which it can be observed that only the associated mutual information of the Yaverage feature increases from 0.0606 (at the 80 Hz sampling rate) to 0.0620 (at the 65 Hz sampling rate). All the mutual information associated with the other features decrease along with the reducing sampling rate. Through the comparison of the experiment results shown in Figs. 8 and 9, it can be inferred that the effect of the sampling rate reduction is more significant on the mutual information associated with the identity recognition than that of the activity recognition, because the mutual information reduction associated with

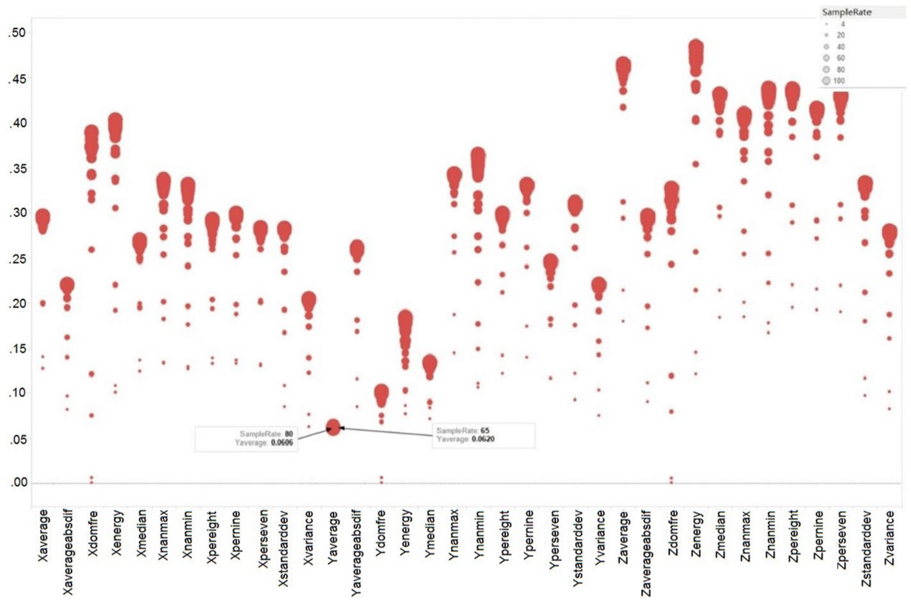


Fig. 9 The variation of privacy information along with the change of the sampling rate

each feature in Fig. 9 is larger than the counterpart in Fig. 8. This observation convinces the contribution of the sampling rate adjustment on the privacy preserving data sharing.

Visualization based privacy-preserving scheme

Three issues exist in the mutual information based scheme proposed in "Mutual information based feature selection and sampling rateadjustment". First, an existing work [14] has shown that both utility and privacy are subjective concepts, which mean that different users (data contributors) may have different preferences. Thus, a practical solution should take into account users' preferences. Second, the mutual information based scheme involves a lot of parameter adjustment, which is hard to use in practice. Finally, it is inconvenient for the mutual information based scheme to identify the appropriate sampling rate for each selected feature.

Design goal and tasks

To address the above issues, in this section, a visualization tool is proposed to interactively analyze the utility and privacy information implied by different combinations of features and sampling rates. The visualization tool still adopts the mutual information based metric to evaluate the importance of a feature at a given sampling rate. Besides that, it utilizes the correlation coefficient to characterize the correlation among features, and allow the tool users (i.e., the data managers) to directly observe the effects of the combination of features and sampling rates to better understand the utility and privacy information implied by the data to be shared. Moreover, by interactively selecting the prefered combination of features and sampling rates, the visualization tool can learning users' preference and provide appropriate recommendation for the users to realize the objective of high utility and low risk of privacy leakage.

In particular, the design objective of the visualization tool can be partitioned into the following 3 tasks: overall distribution (task 1), correlation (task 2), variation trend (task 3).

Task 1 (abbreviated as *T1*) is to facilitate users to observe the utility and privacy information associated with different combinations of features and sampling rates. For example, the visualization tool may provide the user with the features with high activity information at a given sampling rate, the features with less identity information at any sampling rate, or the features with high activity and low identity information at a given sampling rate.

Task 2 (abbreviated as *T2*) is to facilitate users to observe the correlation between the features and activity/identity types, as well as the correlation between the different features. For example, it may provide users with the feature types important to activity/identity, the correlation between features with high identity information, or the similarity between any two features.

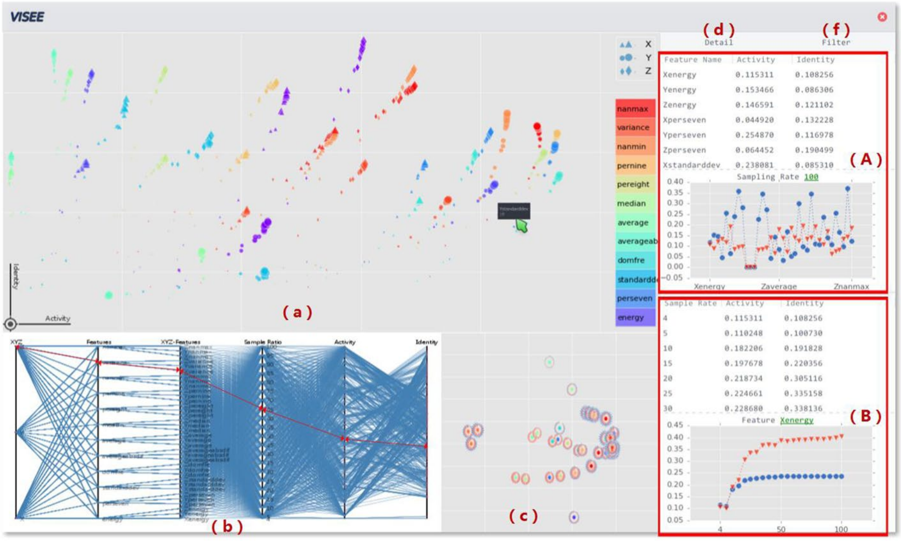
Task 3 (abbreviated as *T3*) is to facilitate users to observe the variation of activity/identity information associated with various features along with the reducing sampling rate. For example, it may provide users the activity information variation along with the reducing sampling rate for a particular feature, the features with small variations within a range of sampling rates, or the features implying even more activity information along with the deduction of sampling rate.

Through realizing the above goals/tasks, the visualization tool proposed in this work provides the opportunities for users to analyze the activity/identity information implied by different combinations of features and sampling rates. Through interactive selection, users can accurately examine the preferred combinations to infer more related information about the data, so that the objective of privacy-preserving data sharing can be realized. The usage of the visualization tool can help users address the following issues in order to figure out a better data sharing solution.

(1) *Feature set selection* Through selecting the preferred features, users can observe the implied utility/privacy information. Through estimating the levels of the implied information, users can infer the corresponding feature types. More importantly, through observing the variation of the utility/privacy information implied by features along with the reducing sampling rate, users can determine the appropriate feature set.

(2) *Sampling rate determination* Through selecting the preferred features, users can observe the variation of the implied information along with the reducing sampling rate. Users can also observe the utility/privacy information implied by different features at a given sampling rate. More importantly, users can determine the appropriate sampling rate by analyzing the common sampling rate, at which features imply high utility information and low privacy information.

(3) *The appropriate combination identification* Through the distribution of the mutual information, users can observe the variation of the utility/privacy information along with the reducing sampling rate. Users can also infer the range of the sampling rate for the preferred features based on the level of the implied utility/privacy information. More importantly, users can also identify the appropriate combination of features and sampling rate by locating the sampling rates, at which given features may imply more/less utility/privacy information.



Feature Name	Activity	Identity
Xenergy	0.115311	0.108256
Yenergy	0.153466	0.086306
Zenergy	0.146591	0.121102
Xperseven	0.044920	0.132228
Yperseven	0.254870	0.116978
Zperseven	0.064452	0.190499
Xstandarddev	0.238081	0.085310

Sample Rate	Activity	Identity
4	0.115311	0.108256
5	0.110248	0.100730
10	0.182206	0.191828
15	0.197678	0.220356
20	0.218734	0.305116
25	0.224661	0.335158
30	0.228680	0.338136

Fig. 10 The interface of the visualization tool, VISEE

The visualization tool design

The interface of the visualization tool, called VISEE, proposed in this work, is shown in Fig. 10, which includes five modules: (a) mutual information distribution diagram, (b) parallel coordinate map, (c) feature grid diagram, (d) mutual information ranking chart, (f) recommended solutions (modules (d) and (f) within the same switchable optional page). In the following, the functionality of each module will be described in details.

Mutual information distribution diagram

The mutual information distribution diagram is the main diagram of VISEE, the main purpose of which is to show the distribution of the utility and privacy information associated with different combination of features and sampling rates. Figure 10a shows the mutual information between the activity/identity types and the features. In this figure, the horizontal axis represents the mutual information between features and activity types, while the vertical axis denotes the mutual information between features and identities. The three geometry shapes represent the feature types in terms of the 3 accelerometer directions: triangles representing X-axis-related features, circles denoting Y-axis-related features, and diamonds meaning Z-axis-related features. Different colors represent different features types, while the sizes of figures denotes the levels of the sampling rates. For example, the icon pointed by the green mouse refers to the Y-Standarddev feature with 10 Hz sampling rate, and the corresponding activity and identity mutual information being 0.428092 and 0.175462, respectively. When a user moves the mouse over a specific icon, the detailed description of the corresponding feature will be popped up.

If a user selects a contiguous preferred region through the left mouse button, VISEE can zoom in the selected region and provide detailed information. Meanwhile, the modules shown in Fig. 10 b, c, f will be updated accordingly. The right mouse button can also

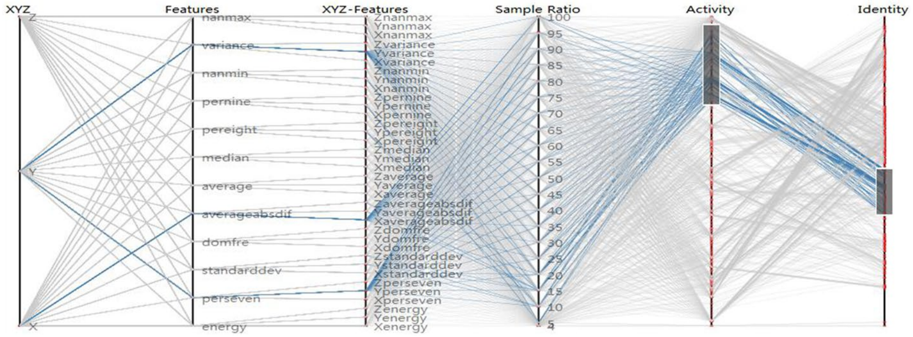


Fig. 11 The interface of the parallel coordinate map

facilitate the user interactive operations by providing a different region selection mode and view restoring.

In a word, users can learn the following through this module: (1) an intuitive understanding to the distribution of mutual information between features and activity/identity types, (2) the trend analysis of the utility/privacy information implied by a particular feature, (3) the observation of the detailed information concerning a particular feature point, (4) the observation and analysis of the preferred feature region. Thus, this module satisfies tasks T1 and T3.

Parallel coordinate map

Parallel coordinate map provides the opportunities for users to select contents based on their preferred values. This map uses a vertical line to represent each factor to be considered, as shown in Fig. 10b, where 6 vertical lines from left to right, represent XYZ axis, Features axis, XYZ-Features axis, Sample Ratio axis, Activity axis, and Identity axis, respectively. The values on each axis are sorted from top to bottom according to certain rules. The three values in XYZ axis and the twelve feature types in Features axis together determines the values in XYZ-Features axis.

A line connecting two axes represent a combination of the values on each individual axis. A polyline traverse the six axes can illustrate six key values. For example, the red polyline shown in Fig. 10b denotes that the mutual information between the X_variance feature and activity/identity types are 0.203750 and 0.202181, respectively, at the sampling rate of 65 Hz.

Users can select the preferred value on any axis to observe the corresponding value range on the other axes. The selected range will be highlighted through a rectangle on the corresponding axis, and the connected lines will be highlighted as blue. It has to be noted that the selection on any axis will influence the highlighted range on other axes, as shown in Fig. 11. The users' interaction on this map will update the modules shown in Fig. 10a, c, f accordingly.

In a summary, users can learn the followings through this module: (1) observing the characteristic of the features and sampling rates associated with the range of selected utility or privacy information, (2) observing the mutual information values based on the selected features, (3) observing the activity/privacy type information implied by features

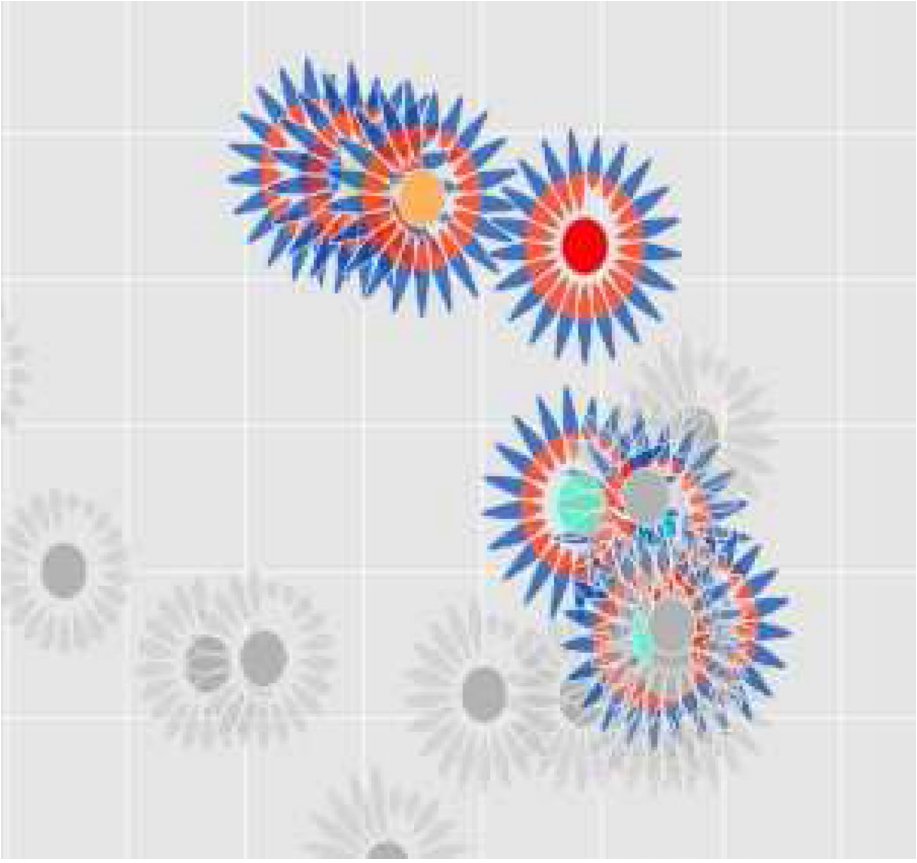


Fig. 12 The interface of the feature grid module

at a particular sampling rate, (4) combining factors to observe the other factors. Thus, this module satisfies tasks T1, T2, and T3.

Feature grid diagram

The feature grid diagram intuitively illustrates the relatively importance of the 36 features and their correlations. As shown in Fig. 10c, each feature is represented as a circle, the center color of which denotes the type of the corresponding feature. Note that the color and feature type mapping is described in Fig. 10a. The whole circle is divided into 21 sectors, each of which corresponds to a specific sampling rate. Each sector associates with two humps. The size of the inner hump in red denotes the strength of privacy information associated with the feature at the corresponding sampling rate, while the size of the outer hump in blue represents the strength of the corresponding utility information, as shown in Fig. 12. When a user move a mouse over any hump, the corresponding detailed description will be popped up, including the corresponding sampling rate and mutual information. The feature grid diagram adopts the MDS embedding technique to visualize the correlation among features. The closer two features in the grid, the more similar they are. This can also help address the feature optimization issues in activity recognition and identity recognition. A user's interaction with this module will update the modules shown in Fig. 10a, b, f accordingly.

In a word, a user can learn the followings through this module: (1) intuitively understanding the utility/privacy information implied by each feature, (2) observing the correlations among features, (3) figuring out the set of features, on which the sampling rate has the significant impact, (4) analyzing feature types and clustering according to the mutual information. Thus, this module satisfies tasks T1, T2 and T3.

Mutual information ranking chart

The major task of the mutual information ranking chart is to enumerate the mutual information between features and activity/identity types at a specific sampling rate, and provide interactive ranking functionality. This module consists of two components, components (A) and (B), as illustrated in Fig. 10d. In component (A), users can select a sampling rate to observe the corresponding mutual information between all features and the activity/identity types. In component (B), users can select a feature to observe the correponding mutual information between the selected feature and the activity/identity types at all possible sampling rates.

In component (A), the sampling rate is chosen through a drop-down list. Once the sampling rate is selected by a user, a table that enumerates all features' associated mutual information will be listed, each row of which consists of feature name, the mutual information related to activity, and the mutual information related to identity. The table is also illustrated intuitively through curves on the bottom of the component, where a user can click the feature name to sort the curve based on the corresponding mutual information. The operations on component (B) are similar.

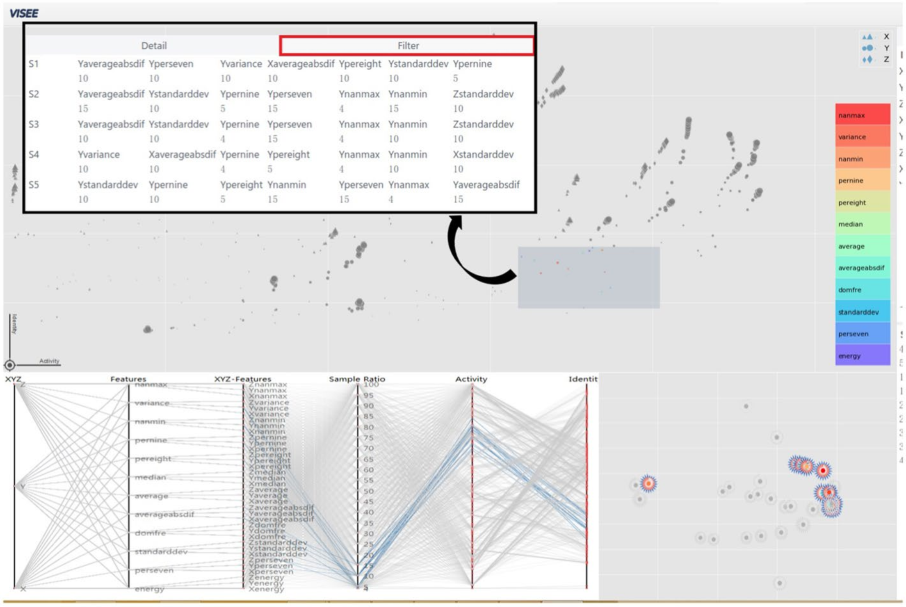
In a word, a user can learn the following through this module: (1) intuitively apprehending the importance of each feature to either utility or privacy, (2) observing the variation of the mutual information associated with a specific feature along with the reducing sampling rate, (3) comparing the utility sensitive features and privacy sensitive features, (4) observing the change of feature types according to the variation of mutual information. Thus, this module satisfies tasks T1 and T3.

Recommended solutions

This module recommends the solution satisfying the objective of high utility and low privacy according to the preferred region selected by a user, as shown in Fig. 13.

To obtain the recommended solution, a user can select a continuous region in the mutual information distribution module. The selected region usually includes a set of features and a set of the corresponding sampling rates. An extended version of the MFS algorithm proposed in "Mutual information based feature selection" is used to determine the appropriate sampling rate for each feature in the selected region. The larger the selected region, the more combinations of features and sampling rates, which in turn incurs more recommended solutions.

This module provides to users only the top combinations in terms of texts, according to the utility and privacy information implied by different combinations. Each combination includes the feature name and its assigned sampling rate. The solution recommended by



	Detail			Filter			
S1	Yaverageabsdif 10	Yperseven 10	Yvariance 10	Xaverageabsdif 10	Ypereight 10	Ystandarddev 10	Ypernine 5
S2	Yaverageabsdif 15	Ystandarddev 10	Ypernine 5	Yperseven 15	Ynanmax 4	Ynanmin 15	Zstandarddev 10
S3	Yaverageabsdif 10	Ystandarddev 10	Ypernine 4	Yperseven 15	Ynanmax 4	Ynanmin 10	Zstandarddev 10
S4	Yvariance 10	Xaverageabsdif 10	Ypernine 4	Ypereight 5	Ynanmax 4	Ynanmin 10	Xstandarddev 10
S5	Ystandarddev 10	Ypernine 10	Ypereight 5	Ynanmin 15	Yperseven 15	Ynanmax 4	Yaverageabsdif 15

Fig. 13 The combination of features and sampling rates recommended by VISEE

this module is only a reference solution. Users can customize their solutions based on the observations on the other modules through comparison.

Evaluation through case study

The privacy preserving scheme through VISEE

To show the effectiveness of VISEE, we first illustrated it through using VISEE to select the appropriate combination of features and sampling rates. To satisfy the objective of high utility and low privacy, a user may select a region in the mutual information distribution diagram with a high activity related mutual information range and low identity related mutual information range, as shown in the gray rectangle of Fig. 13. Based on the user's preference, VISEE computed the recommended combination of features and sampling rates, as shown in the table pointed by the arrow in Fig. 13. From the parallel coordinate map, it can be observed that the computed combination implies high utility information and low privacy information, as shown in the instance of the parallel coordinate map at the bottom left corner of Fig. 13.

To verify the effectiveness of the 5 solutions recommended by VISEE, we conducted an experiment on the accuracies of activity recognition and identity recognition through various learning models mentioned in "Factor Analysis for the Utility and Privacy of AccelerometerData Sharing". The experiment results are shown in Table 5, from which it can be concluded that the recommended solutions can reduce the accuracy of the identity recognition below 18% and improve the accuracy of the activity recognition above 67%, with the discrepancy around 55%, which is much better than the previous learning model based solutions and the mutual information based solution.

Table 5 The accuracies of activity and identity recognitions based on the recommended solutions by VISEE

Recommended	RF activity (%)	RF identity (%)	LMT identity (%)	MP identity (%)
S1	68.4253	12.5119	14.1339	15.6517
S2	70.3982	14.8942	13.955	16.0516
S3	66.8203	8.4337	10.8623	11.8205
S4	67.5857	11.7731	11.7256	13.4807
S5	67.7499	9.4109	9.6575	17.892

Underlined values indicate the best accuracy result achieved by a certain model among all available models

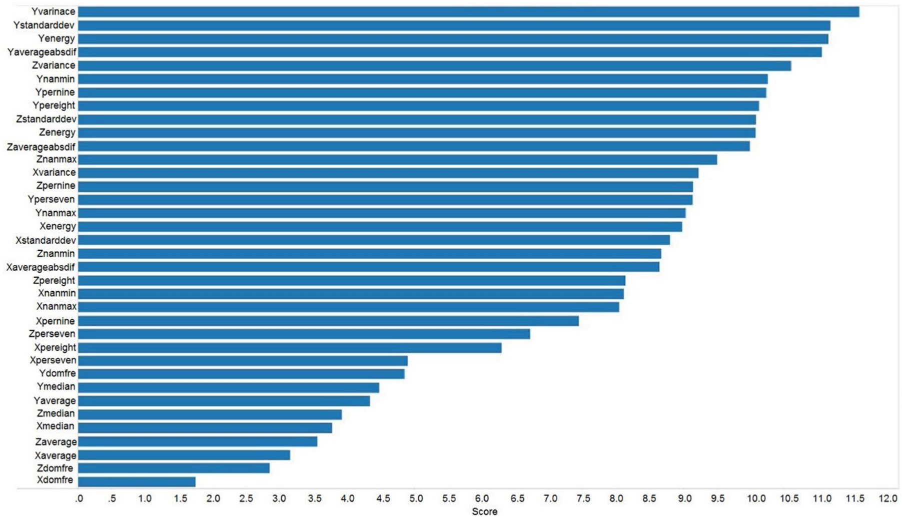


Fig. 14 The feature ranking associated with the utility information

Evaluation on the other data set

To verify that the proposed scheme can apply to the data set other than the HASC data set, we collected smartphone accelerometer data from 20 volunteers. The activity types include walking, running, riding, and standing. To exclude the influence of devices, all volunteers used the same smartphone for the experiments, with the sampling rate being 100 Hz. During the experiments, all volunteers are required to wear tight jeans with the smartphone put into the front pocket, where the three-axis directions are consistent to those of the HASC data set. The experiment location is the sport playground. For each activity type, the range of data collecting time for each volunteer is from 100 to 300s.

Based on the MFS algorithm proposed in "Mutual information based feature selection and sampling rateadjustment", the feature rankings according to the utility information and privacy information can be calculated, as shown in Figs. 14 and 15, respectively. From Fig. 14, it can be observed that, for the utility information, the top 10 features is F_4 ={Yvariance, Ystandarddev, Yenergy, Yaverageabsdif, Zvariance, Ynanmin, Ypernine, Ypereight, Zstandarddev, Zenergy}, while from Fig. 15, we can obtain the top 10 privacy information related features: F_5 ={Zvariance, Zenergy, Yvariance, Xvariance, Xenergy, Znanmin, Zaverageabsdif, Zstandarddev, Xnanmin, Znanmax}. Therefore, for the utility information, Y axis related features have more contributions, while for the privacy

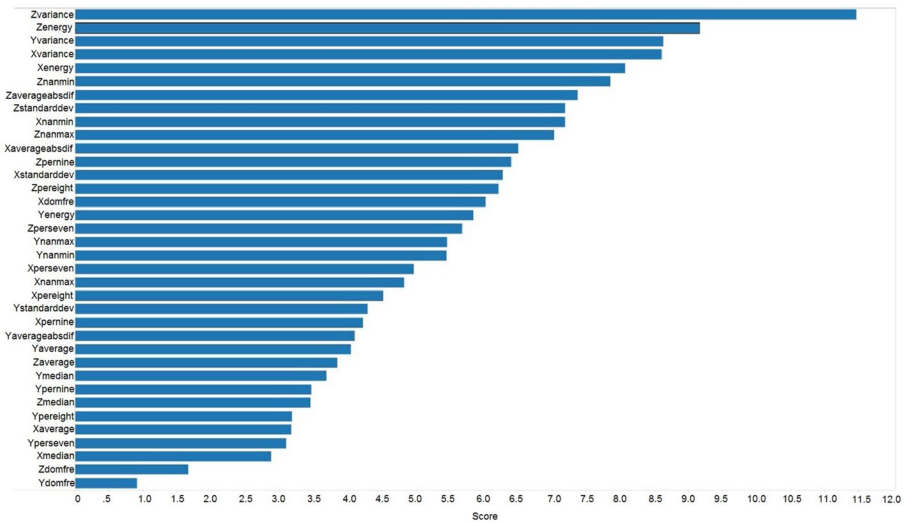


Fig. 15 The feature ranking associated with the privacy information

Table 6 The accuracies of activity and identity recognitions based on feature set F_6

Type	RF activity (%)	RF identity (%)	LMT identity (%)	MP identity (%)
Accuracy	96.646	59.638	64.203	61.159

Underlined values indicate the best accuracy result achieved by a certain model among all available models

information, Z axis related features have more contributions. These are consistent with those of the HASC data set described in "Mutual information based feature selection and sampling rateadjustment".

Then, through the MFS algorithm for feature selection, the feature set F_6 ={Ypereight, Ypernine, Yaverageabsdif, Ystandarddev, Yperseven, Yannmin, Yenergy, Ydomfre} was selected. Most of the features are related to Y axis. The selected feature set is used to evaluate the activity recognition and identity recognition accuracies, as shown in Table 6, where the accuracy difference between the activity recognition and the identity recognition is up to 32.443%.

For the sampling rate adjustment, we also observed the variation of the utility and privacy information along with the adjustment of the sampling rate. The sampling rate was from 100 to 4 Hz, which is consistent with the previous setting for HASC data set. For the learning models, the RF model was used for activity recognition, while the RF, LMT, and MP models were used for identity recognition. The experiment result is illustrated in Fig. 16, from which it can be observed that the accuracy of the activity recognition decreases significantly when the sampling rate reduced from 45 to 4 Hz (Fig. 16a), but the reduction of the identity recognitions accuracy is much more than that of the activity recognition (Fig. 16b), with the difference being 41.0172%.

Finally, VISEE was used to select the combinations of features and sampling rates, and the selected combination was used to compare the accuracy difference between the activity recognition and identity recognition. The result is shown in Table 7, from which

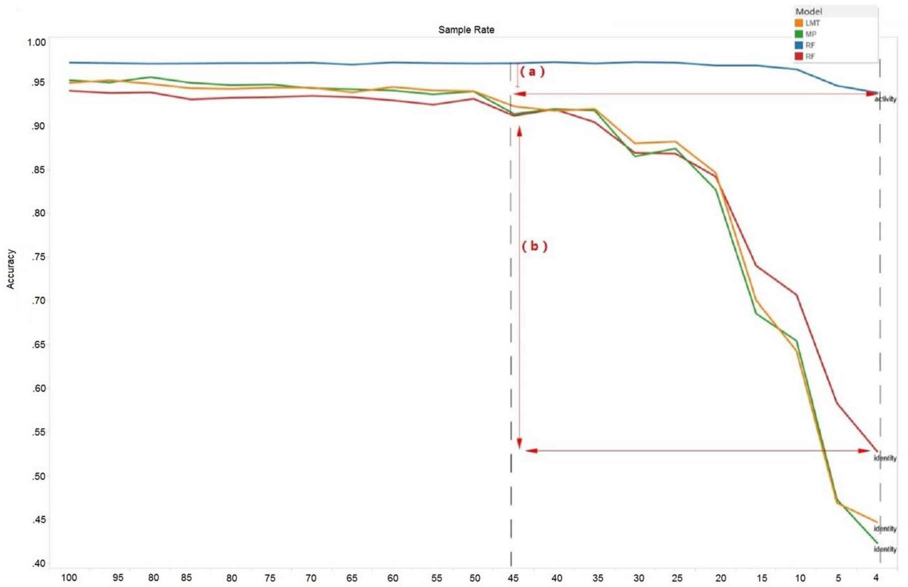


Fig. 16 Accuracy VS sampling rate

Table 7 The accuracies of activity and identity recognitions through the recommended combination from VISEE

Type	RF activity (%)	RF identity (%)	LMT identity (%)	MP identity (%)
Accuracy	95.1765	32.4316	29.6331	29.6331

Underlined values indicate the best accuracy result achieved by a certain model among all available models

it can be observed that the accuracy difference is up to 62.7749%. This result is much better than the results shown in Table 6 and Fig. 16. The experiment results not only convinced us that the effectiveness of the mutual information based feature selection and sampling rate adjustment, but also verified the robustness of VISEE.

Extension and discusssion

Previously, we considered the activity recognition as the application utility and the identity recognition as the privacy to be protected. In this section, we consider more data attributes, such as gender, height, weight, and age, as potential utility or privacy information. To recognized the attributes of gender, height, weight, and age, the acceleration data of 100 users from the HASC data set were preprocessed and extracted features as those described in "Factor analysis for the utility and privacy of accelerometerdata sharing". Then the RF model was applied to recognize gender, height, weight, and age, respectively.

To better fit the RF model, the height attributes of the data set is clustered as follows: 150cm (146–165 cm), 160 cm (156–165 cm), 170 cm (166–175 cm), 180 cm (176–185 cm), and 190 cm (186–195 cm), while the weight attributes is clustered similarly as follows: 40 Kg (45–55 Kg), 50 Kg (56–65 Kg), 60 Kg (66–75 Kg), 70 Kg (76–85 Kg), 80 Kg (86–95 Kg), and 90 Kg (96–105 Kg). The age attribute is used through the information

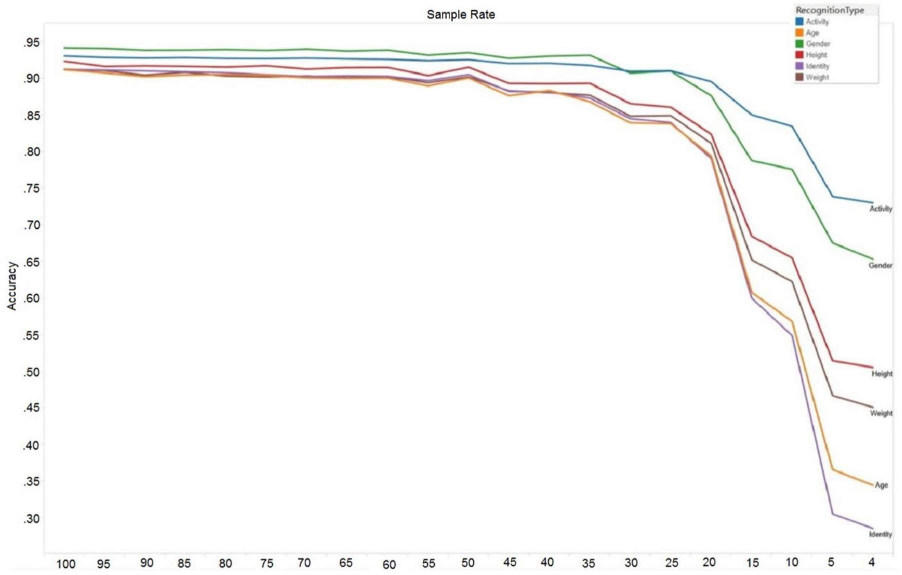


Fig. 17 The impact of sampling rate on the recognition accuracy of all the utility and privacy attributes

provided in the data set, such as “20, early” (age range from 20 to 25 years old) and “20, late” (age range from 26 to 30).

To observe the accuracy reduction of the RF model on all the 6 types of attributes along with the reducing sampling rate, the sampling rate was adjusted from 100 to 4 Hz. The experiment results is shown in Fig. 17, from which it can be observed that the recognition accuracy of all the other attributes reduce faster than that of the activity recognition. It can also be observed that the identity attribute is the most sensitive to sampling rate reduction, while the activity attribute is the least. For gender recognition, it only has two classes (male and female), and it is more sensitive to sampling rate reduction. This implies that the number of patterns to be recognized does not affect the recognition accuracy. Figure 17 also shows the potential privacy protection design space, as the privacy attributes, such as gender, may also be regarded as the application attributes. In that case, if the identity attribute is adopted as the privacy to be protected, it is possible to achieve a good balance between privacy and utility. However, if the height attribute is chosen as the application attribute, while the gender attribute is chosen as the privacy attribute to be protected, it is impossible to protect the privacy and satisfy the application utility at the same time.

To observe the impact of features on the recognition accuracy of those attributes, similar experiments as those in “Mutual information based feature selection and sampling rateadjustment” were conducted. The experiment results are shown in Fig. 18, from which it could be inferred that the features contributing the most to the attributes other than activity are Z axis related features, while the Y axis related features contributed the most to the activity attribute. These results are consistent with those described in previous sections. Moreover, there exists a inclusion relation among identity, age, height, weight, and gender attributes, in the order from left to right, with the privacy protection of the previous one implies the privacy protection for the next one. This also illustrates

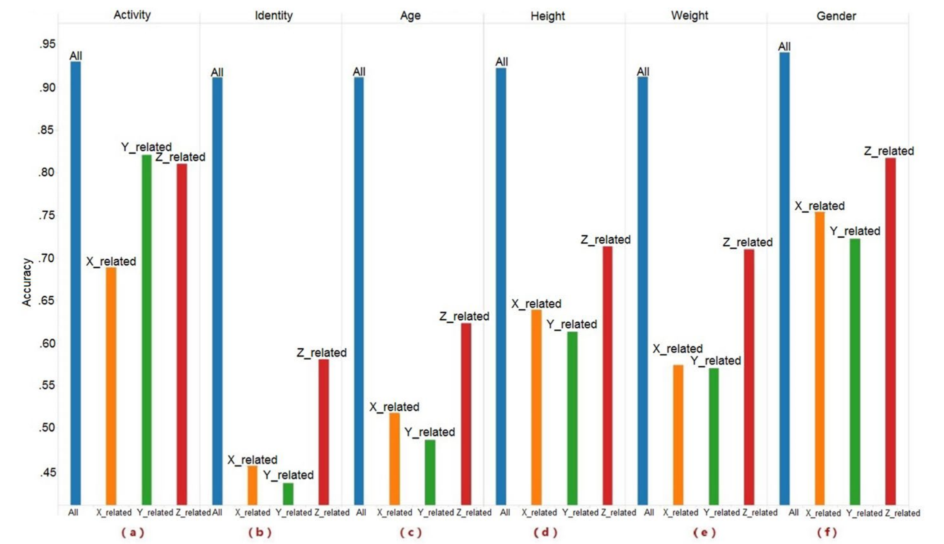


Fig. 18 The impact of features on the other privacy

the possible design space for the information-aware privacy preserving sensing data sharing, considering the trade-off between utility and privacy, which may worth further investigations.

Conclusion

This work proposed a privacy-preserving sensing data sharing solution, which can balance the application utility from data consumers' requests and the privacy concerns from the data contributors. Based on an explanatory sensing data, the accelerometer data, which is widely available across billions of mobile devices, this work proposed a mutual information based feature selection and sampling rate adjustment scheme, which is further extended to an interactive visualization tool to include users' preferences into the solution design space. Intensive experiments have illustrated the effectiveness of the proposed solution. In the future, the authors will generalize this work to include more application scenarios, utilize inferential privacy models to provide mathematical guarantee for our work, and adopt deep learning to enable automatic feature generating.

Authors' contributions

DM, YZ, SM searched the literature; SM, JL, SX, and FXo drew the figures; FX, DM, ML, and YZ conceived and designed the whole system; FX, ML, SM, and SX collected the data; DM, JL, YZ, and SM analyzed the data; JL, YZ, and FX design the visualization tool; FX, DM, JL, ML, and YZ made the data interpretation; FX, DM, YZ, and ML conceived and designed the experiments; SX and SM performed the experiments; FX, ML, SX, SM, CL, and YZ wrote the paper. All authors read and approved the final mansucript.

Author details

[1] School of Humanities and Social Sciences, Beihang University, Beijing, China. [2] Hangzhou Dianzi University, Hangzhou, China. [3] School of Information Science and Engineering, Central South University, Changsha, China.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The data used in this paper can be downloaded from the following links:http://hasc.jp/hc2010/HASC2010corpus/hasc2010corpus-en.html

Ethics approval and consent to participate

Written informed consent was obtained from the research participants for the publication of this report and any accompanying images.

Funding

This work was partially supported by Project Nos. 61232001, 61173169,91646115, and 60903222 supported by National Science Foundation ofChina, No. 2016JJ2149 supported by Science Foundation of Hunan, and alsosupported by the Major Science & Technology Research Program for StrategicEmerging Industry of Hunan (Grant No. 2012GK4054).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 26 January 2018 Accepted: 24 April 2018
Published online: 14 May 2018

References

1. Kawaguchi N, Ogawa N, Iwasaki Y, Kaji K, Terada T, Murao K, Inoue S, Kawahara Y, Sumi Y, Nishio N (2011) HASC Challenge: gathering large scale human activity corpus for the real-world activity understandings. In: Proceedings of augmented human international conference, AH, pp 1–5
2. Ngo TT, Makihara Y, Nagahara H, Mukaigawa Y, Yagi Y (2014) The largest inertial sensor-based gait database and performance evaluation of gait-based personal authentication. Pattern Recognit 47:228–237
3. Wagner DT, Rice A, Beresford AR (2013) Device analyzer: understanding smartphone usage. In: International conference on mobile and ubiquitous systems: computing, networking and services, pp 195–208
4. Favela J, Castro LA, Michan L. Towards a federated repository of mobile sensing datasets for pervasive healthcare. In: Proceedings of the EAI international conference on pervasive computing technologies for kealthcare
5. Song H, Srinivasan R, Sookoor T, Jeschke S, Chowdhury C, Roy S (2017) Mobile crowd sensing for Smart Cities. Smart Cities. John Wiley & Sons, Inc, Hoboken, pp 125–154
6. Triantafyllidis A, Velardo C, Salvi D, Shah SA, Koutkias V, Tarassenko L (2015) A survey of mobile phone sensing, self-reporting and social sharing for pervasive healthcare. IEEE J Biomed Health Inf 21(1):218
7. Chen Y, Xue Y (2016) A deep learning approach to human activity recognition based on single accelerometer. In: IEEE international conference on systems, man, and cybernetics, pp 1488–1492
8. Fung BCM, Wang K, Chen RYuPS (2010) Privacy-preserving data publishing. ACM comput surveys 42(4):14
9. Boldyreva A, Chenette N, Lee Y, ONeill A (2009) Order-preserving symmetric encryption. Advances in cryptology—EUROCRYPT 2009. In: Proceedings of international conference on the theory and applications of cryptographic techniques. 5479:224–241
10. Li N. Li T. Venkatasubramanian S (2007) t-Closeness: privacy beyond k-anonymity and l-diversity. In: IEEE international conference on data engineering, pp 106–115
11. Dwork C (2008) Differential privacy: a survey of results. In: proceedings of the international conference on theory and applications of models of computation, pp 1–19
12. Wang K, Wang P, Fu AW, Wong CW (2012) Inferential or differential: privacy laws dictate. eprint Arxiv, abs/1202.3686
13. Ghosh A, Kleinberg R (2017) Inferential privacy guarantees for differentially private mechanisms. eprint Arxiv, abs/1603.01508
14. Lin J (2013) Understanding and capturing people's mobile app privacy preferences. Dissertations and Theses—Gradworks
15. Kwapisz JR, Weiss GM, Moore SA (2011) Activity recognition using cell phone accelerometers. ACM SigKDD Explor Newslett 12(2):74–82
16. Adam NR, Worthmann JC (1989) Security-control methods for statistical databases: a comparative study. ACM Comput Surveys 21(4):515–556
17. Agrawal R, Srikant R (2000) Privacy-preserving data mining. In: ACM SIGMOD international conference on mof data, pp 439–450
18. Liu Q, Wang G, Li F, Yang S, Wu J (2017) Preserving privacy with probabilistic indistinguishability in weighted social networks. IEEE Trans Parallel Distrib Syst 28(5):1417–1429
19. Luo E, Liu Q, Abawajy JH, Wang G (2017) Privacy-preserving multi-hop profile-matching protocol for proximity mobile social networks. Future Gener Comput Syst 68:222–223
20. Gao C, Cheng Q, He P, Susilo W, Li J (2018) Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack. Inf Sci 444:72–88
21. Peng T, Liu Q, Meng D, Wang G (2017) Collaborative trajectory privacy preserving scheme in location-based services. Inf Sci 387:165–179
22. Kumari V, Chakravarthy S (2016) Cooperative privacy game: a novel strategy for preserving privacy in data publishing. Humancentric Comput Inf Sci 6(1):12
23. Blundo C, Orciuoli F, Parente M (2017) An Am I-based and privacy-preserving shopping mall model. Humancentric Comput Inf Sci 7(1):26
24. Gai K, Qiu M, Zhao H (2017) Privacy-preserving data encryption strategy for big data in mobile cloud computing. IEEE Trans Big Data 1. https://doi.org/10.1109/TBDATA.2017.2705807
25. Chen F, Wang S, Jiang X, Ding S, Lu Y, Kim J, Sahinalp SC, Shimizu C, Burns JC, Wright VJ (2017) PRINCESS: privacy-protecting rare disease international network collaboration via encryption through software guard extensions. Bioinformatics 33(6):871

26. Luo E, Liu Q, Wang G (2016) Hierarchical multi-authority and attribute-based encryption friend discovery scheme in mobile social networks. IEEE Commun Lett 20(9):1772–1775
27. Gao C, Cheng Q, Li X, Xia S. Cloud-assisted privacy-preserving profile-matching scheme under multiple keys in mobile social network. Cluster Comput 2018. https://doi.org/10.1007/s10586-017-1649-y
28. Li P, Li J, Huang Z, Li T, Gao C, Yiu S, Chen K (2017) Multi-key privacy-preserving deep learning in cloud computing. Future Gener Comput Syst 74:76–85
29. Zhu T, Zou X, Pan J (2017) Query with SUM aggregate function on encrypted floating-point numbers in cloud. J Inf Process Syst 3(13):573–589
30. Van NB, Lee S, Kwon K (2017) Selective encryption algorithm using hybrid transform for GIS vector map. J Inf Process Syst 13(1):68–82
31. Sweeney L (2002) k-ANONYMITY: A Model for Protecting Privacy. Int J Uncertain Fuzziness KnowledgeBased Syst 10(5):557–570
32. Machanavajjhala A, Gehrke J, Kifer D, Venkitasubramaniam M (2006) L-diversity: privacy beyond k-anonymity. In: Proceedings of the international conference on data engineering, pp 24–24
33. Perentis C, Vescovi M, Lepri B (2015) Investigating factors affecting personal data disclosure. In: Proceedings of the international conference on world wide web, pp 89–90
34. Perentis C, Vescovi M, Leonardi C, Moiso C, Musolesi M, Pianesi F, Lepri B (2017) Anonymous or not? Understanding the factors affecting personal mobile data disclosure. ACM Trans Internet Technol 17(2):13
35. Guo B, Nixon MS (2008) Gait feature subset selection by mutual information. IEEE Trans Syst Man Cybern Part A Syst Humans 39(1):36–46
36. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27(8):1226–1238
37. Lu M, Guo Y, Meng D, Li C, Zhao Y (2017) An information-aware privacy-preserving accelerometer data sharing. In: International conference of pioneering computer scientists, engineers and educators
38. Incel OD, Kose M, Ersoy C (2013) A review and taxonomy of activity recognition on mobile phones. Bionanoscience 3(2):145–171
39. Kwapisz JR, Weiss GM, Moore SA (2010) Cell phone-based biometric identification. In: Proceedings of the IEEE international conference on biometrics: theory applications and Systems, pp 1–7
40. Derawi M, Bours P (2013) Gait and activity recognition using commercial phones. Comput Secur 39:137–144
41. Shoaib M, Bosch S, Incel OD, Scholten H, Havinga PJ (2015) A survey of online activity recognition using mobile phones. Sensors 15(1):2059–2085
42. Ailisto HJ, Makela SM (2005) Identifying people from gait pattern with accelerometers. Proc SPIE Int Soc Opt Eng 5779:7–14
43. Bishop CM (2006) Pattern Recognit Mach Learn. Springer, New York, Inc, Information Science and Statistics, Berlin
44. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. IEEE Trans Neural Netw 5(4):537–550
45. Bassir SM, Akbari A, Nassersharif B (2014) An improved feature transformation method using mutual information. Int J Speech Technol 17(2):107–115