Privacy preserving data publishing of categorical data through *k*-anonymity and feature selection

Aristos Aristodimou ✉, *Athos Antoniades, Constantinos S. Pattichis*

Department of Computer Science, University of Cyprus, Nicosia, Cyprus
✉ *E-mail: aristodimou.aristos@ucy.ac.cy*

Published in Healthcare Technology Letters; Received on 18th November 2015; Revised on 7th March 2016; Accepted on 9th March 2016

In healthcare, there is a vast amount of patients' data, which can lead to important discoveries if combined. Due to legal and ethical issues, such data cannot be shared and hence such information is underused. A new area of research has emerged, called privacy preserving data publishing (PPDP), which aims in sharing data in a way that privacy is preserved while the information lost is kept at a minimum. In this Letter, a new anonymisation algorithm for PPDP is proposed, which is based on *k*-anonymity through pattern-based multidimensional suppression (kPB-MS). The algorithm uses feature selection for reducing the data dimensionality and then combines attribute and record suppression for obtaining *k*-anonymity. Five datasets from different areas of life sciences [RETINOPATHY, Single Proton Emission Computed Tomography imaging, gene sequencing and drug discovery (two datasets)], were anonymised with kPB-MS. The produced anonymised datasets were evaluated using four different classifiers and in 74% of the test cases, they produced similar or better accuracies than using the full datasets.

Nomenclature

x	single value variable
$\boldsymbol{x}$	vector
$\|\boldsymbol{x}\|$	number of elements in vector $\boldsymbol{x}$
X	$n \times d$ matrix
$X_{i,\,j}$	value of the jth feature of the ith instance

1. Introduction: In recent years, with the infiltration of information technology in healthcare, many healthcare related entities such as hospitals and pharmaceutical companies, have vast amounts of patients' data. Although it is clear that sharing such data can increase the likelihood of identifying novel findings or even replicating existing research results, this is not happening due to legal and ethical issues. Attempts have been made by research programs like Linked2Safety [1], for merging data across different healthcare entities and preserve individuals privacy, but one needs to also decide which data to share, so that the likelihood that the shared data will lead to new findings is increased. The problem is more evident in the case of genetic data, since they can reveal the identity of an individual and possible risk factors that she/he may have. Hence, they are highly sensitive data and caution is needed even if a part of them is shared.

The aim of privacy-preserving data publishing (PPDP) is to provide the means for publishing data in a way so that the privacy of individuals is preserved with a minimum loss of information [2]. One approach that is employed for data anonymisation is *k*-anonymity [3, 4]. In *k*-anonymity, a dataset is considered anonymised if the combined values of the *quasi-identifiers* appear at least k times, which means that there is at most $1/k$ probability of identifying an individual using the available data. Features are considered as *quasi-identifiers*, if their combined values can be linked to publicly available information and can lead to the re-identification of individuals [4]. Knowing beforehand all of the features that can be used as *quasi-identifiers* are difficult [4] (e.g. genetic data), hence one could consider all features as such and therefore publish datasets whose records exist at least k times.

The most common methods for achieving the *k*-anonymity requirement are generalisation and suppression [5]. In generalisation, the values of features are transformed into more general ones so that details of the individuals are disclosed. For example, the address of an individual could be replaced with the zip code, the city or even the country the individual is living, depending on how abstract the information needs to be so that *k*-anonymity is obtained. This requires the creation of value generalisation hierarchies (taxonomy trees) for the *quasi-identifiers*, so that an anonymisation algorithm can select the minimum level of generalisation needed. Research focused more on the generalisation approach, resulting in many such algorithms [6–9], but their need for taxonomy trees makes it hard to use in high-dimensional data. Algorithms that can cope with the lack of user defined taxonomy trees such as TDR [10] exist, but in general, high-dimensionality can severely affect the information contained in the anonymised datasets [11].

In suppression, part of the dataset is removed so that the *k*-anonymity requirement is not violated. This can be achieved by various suppression methods, such as record suppression and attribute suppression. For example, in record suppression, all records (instances) that appear less than k times in the dataset are removed, whereas in attribute suppression, features that have values that do not conform to *k*-anonymity are removed. This approach can lead to substantial information loss since data are removed instead of being replaced by generalised values, hence caution is needed when used. Methods that employ suppression for obtaining *k*-anonymity before data publishing have been proposed [4, 12, 13] and there has been some focus on suppression methods that also combine feature selection [14, 15]. With the use of feature selection, the dimensionality of the data is reduced and features that are not related to the problem of interest can be removed. With fewer features, the probability of having records that are unique decreases, thus less suppression is required for obtaining *k*-anonymity.

In this paper, *k*-anonymity through pattern-based multidimensional suppression (kPB-MS) is proposed for PPDP. The algorithm uses feature selection for reducing the data dimensionality and then combines attribute and record suppression for obtaining *k*-anonymity. The proposed algorithm can be used on categorical data for classification tasks. The remaining of this Letter is organised as follows: Section 2 introduces a new measure, which is used in the proposed feature selection algorithm presented in Section 3. In Section 4, the anonymisation methodology is

doi: 10.1049/htl.2015.0050

Table 1 Example of a contingency table

f_1	f_2	f_3	Class = 0	Class = 1
1	2	1	10	0
1	2	2	0	10
2	3	1	10	20
2	3	3	10	10
3	1	1	20	0
3	1	3	0	10

shown, whereas Section 5 presents the datasets and the evaluation methodology used. The last three sections contain the results of the algorithm along with a discussion and the concluding remarks.

2. Pattern-based classification accuracy (PBCA): In this section, a new measure is proposed, which will be used as the performance metric of the feature selection step of the anonymisation methodology. PBCA uses the discrimination power of the patterns in a dataset to calculate the accuracy that can be obtained by classifiers. It is applicable on categorical data, and considers each input instance as a pattern. As will be shown it is the upper limit of the classification accuracy that can be obtained by classifiers when the entire dataset is used both for training and testing. The mathematical notations that will be used in this Letter are given in the Nomenclature section.

The PBCA of the features in X and the response variable $\boldsymbol{y}$, can be calculated using the following equation

$$\text{PBCA}(X, \boldsymbol{y}) = \frac{\sum_{i=1}^{p} \max(T_{i,c} : c \in [1, |\boldsymbol{c}|])}{n} \tag{1}$$

where p is the number of rows of the contingency table (number of unique patterns), T is the contingency table, $T_{i,c}$ is the number of instances of row/pattern i in the contingency table T that belong to class c, $|\boldsymbol{c}|$ is the number of classes of the response variable and n is the number of instances in the dataset.

PBCA is model free and can capture both linear and non-linear dependencies among the features and the response variable. The complexity of the approach for calculating the PBCA is $O(n)$, since it requires one pass from the data to create the contingency table and a single pass from the contingency table to calculate the PBCA.

A disadvantage of the measure is that it is biased towards variables with many categorical values. For example, if the unique identifier of each instance is included in the dataset, then the PBCA would be 100%, regardless of the values of the rest of the variables.

Table 1 shows a contingency table of a hypothetical dataset. The dataset has three features (f_1, f_2, f_3) with three categorical values each, whereas the response variable (class) is binary. The first row indicates that all ten instances with the values (pattern) {1, 2, 1} for features f_1, f_2 and f_3, respectively, belong to class '0'. Similarly, the rest of the rows indicate how many instances for each pattern belong to each class in the dataset. Such a contingency table can be created with a single pass from the dataset using a hash table, in which the key used is the pattern of the instances.

When applying (1) on the dataset in Table 1, a PBCA of 80% is obtained, which means that $(10 + 10 + 20 + 10 + 20 + 10) = 80$ instances out of the total 100 would be correctly classified.

An ideal classifier would classify an instance to class '0' if most of the instances of a pattern belonged to '0', and '1' otherwise. Using the dataset shown in Table 1, it would classify instances that match the pattern of the first row as class '0', instances that match the pattern of the second row as class '1' and so on. In the case of a tie among the number of instances that belong to each class of a pattern, any of the classes can be selected. Hence by selecting the most probable class for each pattern, as observed in the dataset, the PBCA for the dataset can be calculated. As can be seen, since the most probable class for each pattern is selected, the PBCA indicates the upper limit of the classification accuracy that can be obtained when the entire dataset is used for both training and testing by a classifier.

Algorithm 1

Require: $X, \mathbf{y}$
1: $\mathbf{s} \leftarrow FSFS(X, \mathbf{y})$
2: $\mathbf{s} \leftarrow removeRedundantFeatures(\mathbf{s}, X, \mathbf{y})$
3: $\mathbf{s} \leftarrow sortFeatures(\mathbf{s}, X, \mathbf{y})$
4: **return** $\mathbf{s}$

Fig. 1 *PB-FSS*

3. Pattern-based feature subset selection (PB-FSS): As mentioned, with feature selection one can reduce the dimensionality of a dataset and remove any redundant and non-informative features. For the needs of the proposed anonymisation process, a new feature selection algorithm is presented and shown in Algorithm 1 (see Fig. 1). PB-FSS has three main steps. The first step is to perform forward sequential feature selection (FSFS) using PBCA as its performance metric. The second step is to remove any features that are redundant in the features subset and the final step is to order the selected features based on their importance.

FSFS begins with empty features subsets and sequentially adds the features that increase the performance metric. At each iteration of the FSFS, the feature that maximises PBCA when added to the selected features subset is found. If the selected feature increases the PBCA when included in the features subset $\boldsymbol{s}$, the process is repeated. If the PBCA is not increased then the process finishes and the feature subset without this feature is returned. Since the entire process depends on the PBCA, to overcome its bias towards features with many categorical values, cross-validation is used in the PBCA calculation.

When adding features in the subset using FSFS, it is possible that some of the already selected ones become redundant. For example, a new feature that is selected, interacts with one or more features in a way that the effect of a previously selected feature is masked and hence is no longer needed. In such cases, features that become redundant need to be removed. The second step of PB-FSS is responsible for removing such features and is performed in the function *removeRedundantFeatures*. At each iteration of the function, a feature is removed from the subset and the PBCA is recalculated. If the PBCA remains the same after the removal of a feature, that feature is considered as redundant and it is removed from the selected features subset, otherwise it is kept in the final subset.

The third step of the PB-FSS, sorts the selected features in descending order based on their contribution to the PBCA. To calculate the PBCA contribution of a feature the following procedure is followed. Each feature is removed from the subset to calculate the PBCA that can be obtained without it and then it is added back to the subset. The feature that produced the smallest PBCA difference when removed, is considered as the one with the smallest contribution and is added to the beginning of a new subset list. This is repeated until no features remain in the initial subset list. The reason the sorting is performed in a backward feature selection manner, is that this allows taking into consideration feature interactions and produce a better ranking.

Since PB-FSS is based on the PBCA, which is model free, it can capture any type of non-linear interactions among the features. In addition, the algorithm is able to remove features whose effect can be expressed by the interactions of the other selected features, which can further decrease the size of the returned subset. Finally, it provides a ranking of the features that accounts any

Algorithm 2

Require: $X, \mathbf{y}, k, t$
1: $\mathbf{s} \leftarrow PB\text{-}FSS(X, \mathbf{y})$
2: **while** 1 **do**
3: $Z \leftarrow anonymization(X, \mathbf{y}, \mathbf{s}, k)$
4: $loss \leftarrow instanceLoss(X, Z)$
5: **if** loss > t **then**
6: $\mathbf{s} \leftarrow removeLastFeature(\mathbf{s})$
7: **else**
8: *break*
9: **end if**
10: **end while**
11: **return** Z

Fig. 2 *kPB-MS*

feature interactions that exist. A disadvantage of the method, due to its forward selection strategy, is that it could end up in not selecting interacting features with low main effects, such as in the XOR problem.

4. k-Anonymity through pattern-based multidimensional suppression: Data anonymisation using the suppression method of *k*-anonymity, can result in the removal of a substantial amount of the initial instances, especially when the dataset is high dimensional. In addition, by removing instances from the dataset there is the risk of removing important information, which can lead to poorer classification results or even to wrong models. To address these issues, a new anomysation process called kPB-MS is proposed.

The steps of kPB-MS are shown in Algorithm 2 (see Fig. 2). Initially PB-FSS is performed for reducing the dimensionality of the dataset and for ranking the selected features based on their importance. In the anonymisation step, a contingency table with each pattern in the reduced dataset is created like in Table 1. For each pattern, a new $2 \times |\boldsymbol{c}|$ contingency table is created, where $|\boldsymbol{c}|$ is the number of classes. An example of such a contingency table is shown in Table 2.

The first row has the initial number of instances of each class for that pattern. The second row contains the remaining number of instances once *k*-anonymity is performed. This means that in the second row, the number of instances of the classes that violate *k*-anonymity are replaced with the value of zero. Fisher's exact test is then performed on this contingency table, for testing the significance of the change in the number of instances in each class. In case the change is statistically significant, all of the instances of that pattern are removed, instead of only removing the ones that do not conform to *k*-anonymity.

Finally, the percentage of the lost instances (*loss*) from the anonymisation process is calculated. In case the *loss* is above the user's defined threshold (t), the least important feature (in this case the last feature in the subset) is removed and the process is repeated, otherwise the anonymised dataset Z is returned.

5. Evaluation methodology: In this section, a description of the datasets used is given along with the preprocessing steps applied on them. Then the process followed for the analysis of the datasets is described.

Table 2 Contingency table for testing the effect of *k*-anonymity

	Class = 0	Class = 1
initial	*10*	*20*
k-anonymised ($k = 15$)	*0*	*20*

Table 3 Summary of the datasets

Dataset	Number of features	Number of instances	Number of classes	Class instances
RETINOPATHY	19	1151	2	540/611
SPECT	22	267	2	55/212
SPLICE	60	3190	3	767/768/1655
HTS	1024	3115	2	2000/1115
DOROTHEA	100,000	1150	2	112/1038

5.1. Datasets: A summary of the datasets used in the evaluation is given in Table 3. All of the datasets are from the UCI Machine Learning Repository [16], except HTS which was created from PubChem [http://www.pubchem.ncbi.nlm.nih.gov/assay/assay.cgi?aid=633]. Since the proposed algorithms require categorical data, any continuous features in the datasets were discretised using the minimum description length principle [17].

The datasets are from different areas of life sciences. RETINOPATHY has features extracted from the Messidor image set [18] to predict whether an image contains signs of diabetic retinopathy or not, whereas Single Proton Emission Computed Tomography (SPECT) contains data on cardiac SPECT images and is on the classification of normal and abnormal patients. SPLICE contains gene sequences and is on the identification of splice junctions. SPLICE is the only dataset with three classes. HTS is on drug discovery and was created using the compounds and outcomes from PubChem. The fingerprints were calculated using the LiSIs platform [19] using a radius equal to two and a fingerprint size of 1024. The initial distribution of the classes for this dataset was highly imbalanced (99%/1%), thus resampling was used for selecting a subset of the instances of the frequent class (2000 instances). DOROTHEA is also on drug discovery and it is the largest dataset used in the experiments. This high-dimensional dataset is imbalanced, with its less frequent class representing 11% of the instances.

5.2. Analytical methodology: For the evaluation of kPB-MS, the following methodology was followed. Initially, each dataset was anonymised using kPB-MS. Different k values for *k*-anonymity were used and the accepted loss of instances (t) was set to 10%. Specifically k was set to the values 1, 3, 5, 10 and 20. For the case that k was set to 1, the non-anonymised dataset that results from PB-FSS, is used, so that the effects of k can also be observed. All of the features in the datasets were considered as *quasi-identifiers*.

Once the anonymised datasets were produced, classification with a 10-fold stratified cross-validation was performed on the full and the anonymised datasets. The same folds were used in the training and testing of both the full and the anonymised datasets. Specifically, the test folds were always the same among the full and the anonymised datasets, but the training folds could either be the same or use a subset of the instances in the anonymised datasets. This is due to the fact that instances could be removed during the anonymisation process. The statistical significance among the results was calculated with a paired t-test and it compared the accuracies on the ten test folds of the anonymised and the full datasets. The difference was considered as statistically significant, when the obtained p-value was <0.05.

For the classification task, four non-linear classifiers were selected. The support vector machine (SVM) [20] with its default parameters, the 1-nearest neighbour (1-NN) algorithm [21], the C4.5 decision tree [22] with its default parameters, and the random forest (RF) algorithm [23] using 50 trees. Their accuracies were calculated in *R* using the RWeka package [24] for *k*-NN and C4.5 and the e1071 package for the SVM and the randomForest package [25] for RF.

doi: 10.1049/htl.2015.0050

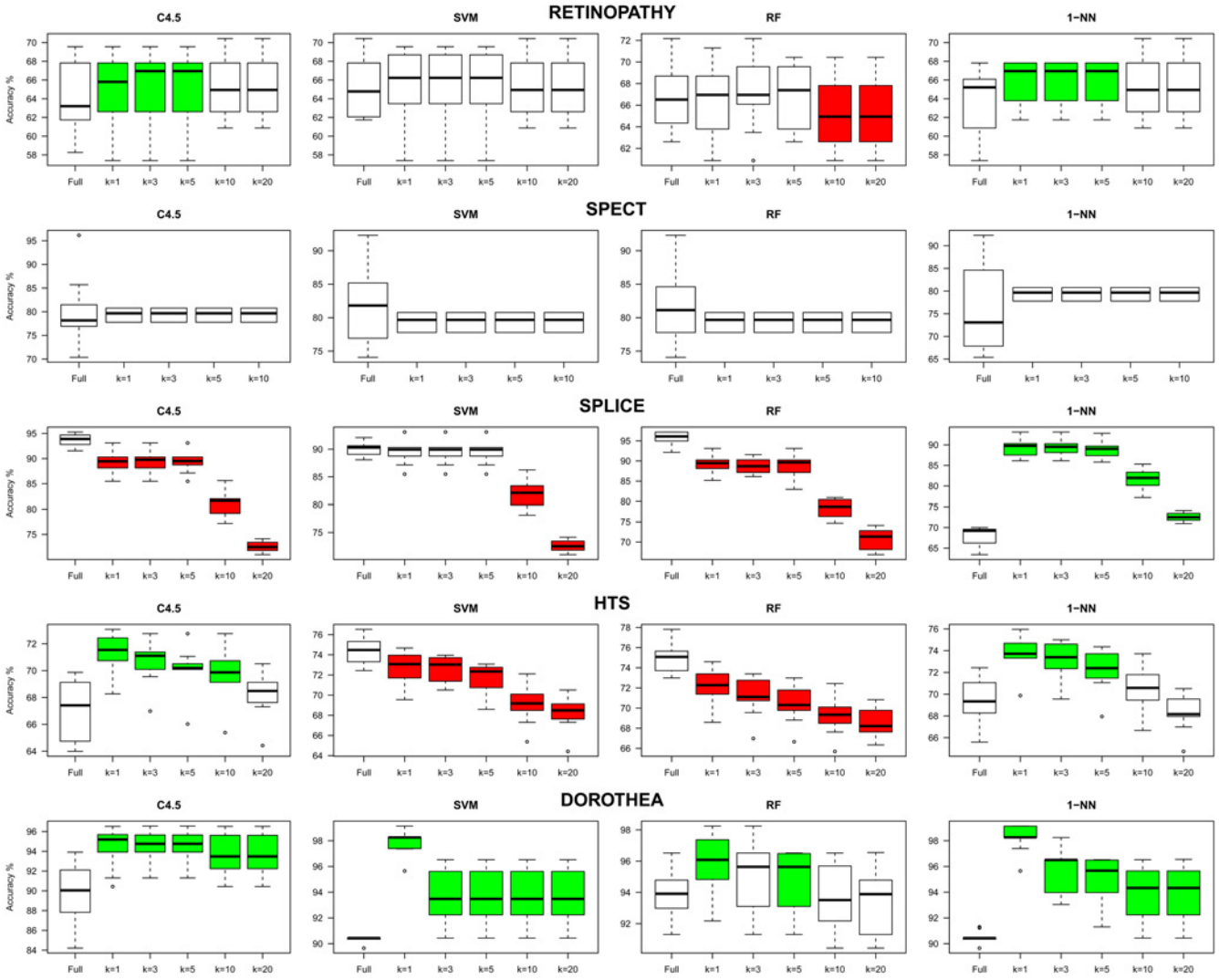

Fig. 3 *Results on the full datasets and the kPB-MS anonymised datasets with different values for k*
Each plot represents the results of a classifier on a specific dataset. The *x*-axis has the values of *k*-anonymity ('Full' means that the full dataset was used) and the *y*-axis has the classification accuracy obtained. Boxplots with a white background indicate that there is no statistical significance (at $p < 0.05$) in the difference among the results of the full dataset and the anonymised dataset. Green boxplots indicate anonymised datasets with better classification results than the full dataset and red boxplots indicate anonymised datasets with worse results than the full dataset

6. Results: For each classifier, its results on the full and anonymised datasets are illustrated in Fig. 3. The *x*-axis of each plot represents the value used in *k*-anonymity, whereas the *y*-axis represents the classification accuracy obtained. The colours of the boxplots indicate if the classification accuracy, using the anonymised dataset had a statistically significant change compared to the one obtained when the full dataset is used. The white colour indicates no significant accuracy differences, the green colour indicates that better accuracies were obtained with the anonymised dataset and the red colour indicates that better accuracies were obtained with the full dataset.

In Table 4, the features retained in the anonymised dataset and the instance loss for different values of k are shown. For the SPECT dataset, with $k = 20$, the removed instances were more than the accepted threshold so no data could be published. For the rest of the datasets, anonymised data could be published for all of the tested values of k.

As can be seen in Fig. 3, the PB-FSS datasets with $k = 1$ did not affect negatively the classifiers accuracies in 80% of the cases. Hence the proposed feature selection algorithm is capable of selecting informative features. With the increase of the value of k, the obtained accuracies are affected as expected [14, 26]. This is due to the fact that more instances are being suppressed, which can cause the use of fewer features when the instances loss is above the pre-specified threshold. The effect is more visible for values larger than five in the datasets tested (see Table 4). This also shows that before publishing data an analysis on the effect of different values of k should be performed before selecting the most appropriate value, so that a balance between privacy and information loss is acquired.

Table 4 The effect of k on the dimensionality and the instance loss

Dataset	Features used (instance loss)				
	$k = 1$	$k = 3$	$k = 5$	$k = 10$	$k = 20$
RETINOPATHY	4 (0%)	4 (3%)	4 (5%)	2 (2%)	2 (5%)
SPECT	1 (0%)	1 (0%)	1 (0%)	1 (0%)	0 (100%)
SPLICE	4 (0%)	4 (2%)	4 (8%)	3 (1%)	2 (6%)
HTS	25 (0%)	25 (7%)	21 (10%)	13 (9%)	8 (9%)
DOROTHEA	19 (0%)	19 (4%)	19 (4%)	19 (8%)	14 (10%)

In general kPB-MS with a $k > 1$ did not have a statistically significant negative effect on the obtained accuracies in 74% of the test cases. The best results were obtained in SPECT and DOROTHEA. In SPECT only one feature was selected and it had enough information to produce similar accuracies with the full dataset. In DOROTHEA, the anonymised datasets had better or similar accuracies with the full dataset since the high dimensionality was affecting the classifiers ability to model the data.

7. Discussion: The proposed kPB-MS, combines feature selection with both attribute and record suppression and additionally provides a method for selecting the accepted percentage of records to be lost. The algorithm attempts to reduce the information lost and as seen from the results it accomplishes this in most of the tested datasets.

The first step of kPB-MS uses the PB-FSS, hence informative features are selected and ranked according to their importance. This can help researchers focus on these features and help them interpret their results. In addition, by sharing the anonymised datasets produced by kPB-MS they increase the probability of replicating the results obtained since they are focusing on informative features.

Since it is using suppression, no user defined taxonomy trees need to be defined as is the case with generalisation techniques [6, 7, 8], which makes the algorithm easier to use by non-domain experts. In addition, one does not need to know beforehand, which features are the possible *quasi-identifiers*, since all of the features can be treated as such. This can further reduce the risk of publishing data that might accidentally contain *quasi-identifiers* that do not adhere to the *k*-anonymity requirement.

kPB-MS uses multidimensional suppression, which means that it is taking into consideration all of the value combinations of the features for preserving *k*-anonymity. A similar approach is also followed by kACTUS [14]. kACTUS uses C4.5 for building a classification tree and uses the selected features of the tree to rank the features and decide which to suppress when they do not comply to *k*-anonymity. Due to the use of C4.5, kACTUS is bound to the performance of the classifier and in high dimensional data like DOROTHEA, kPB-MS is expected to perform better.

In [15], DMPD is proposed, which uses a genetic algorithm for searching for an optimal feature set partitioning. One difference with kPB-MS is that it produces multiple feature subsets, which can all be used by classifiers and then combine their classification predictions. This approach is expected to produce less suppressions than kPB-MS, since it can create many subsets with a small number of features and hence have a smaller probability to not adhere to *k*-anonymity. On the other hand, it might be computationally demanding when used on data which are high dimensional like genetic data, due to the increased search space and the need for parameter optimisation. kPB-MS would require running PB-FSS on the dataset once to get a ranked list of informative features and then the optimisation of k and the instance loss threshold t can be performed on the reduced dataset. Hence it is less computationally demanding since the optimisation of the parameters does not require rerunning the feature selection process on the entire dataset each time. In addition, PB-FSS is also computationally efficient, since it is based on forward sequential selection and its performance metric (PBCA) can be calculated in $O(n)$.

To the authors knowledge, there are no similar studies on the datasets used in this Letter. Both DMPD [15] and kACTUS [14] were shown to have significantly better results than traditional methods, but were not compared with each other. Since the performance of kACTUS is bound to C4.5, it is expected that if evaluated on the five datasets investigated in this study, its performance would be comparable to the C4.5 results documented in Fig. 2 (for the full datasets). Based on this, it can be inferred that kPB-MS is expected to give similar or better results than kACTUS for four out of the five datasets investigated (as documented in Fig. 2 for SPECT, RETINOPATHY, HTS and DOROTHEA).

8. Conclusions: In this Letter, a new method for *k*-anonymising datasets has been proposed. This includes the proposal of a new measure and a new feature selection algorithm along with multidimensional suppression. The proposed measure (PBCA), is model free and has a complexity of $O(n)$, which reduces the computational demands of the feature selection algorithm and makes the whole process applicable for high-dimensional data. With PB-FSS, informative features are selected and ranked so that only such features are shared. This can further increase the probability of sharing data that can lead to replicating results.

The multidimensional suppression procedure of kPB-MS takes into consideration the instances removed and the effect of record suppression on classifiers. The first is obtained by allowing the user to define the accepted loss of instances, whereas the second is obtained by testing the significance of the suppression using Fisher's exact test on each pattern. As shown in the results, the algorithm did not negatively affect the classifiers in 80% of the test cases, indicating that kPB-MS can be used in privacy preserving data publishing.

As seen, the value of k can affect the overall results, hence before publishing the data, a proper value for k should be selected. The selection should be made in a way that it provides a balance between the obtained privacy and the information lost. For this task, the evaluation methodology used in this Letter can be used on different values of k for guiding such a decision.

Since PB-FSS is using a forward selection method, it can miss interacting features with low main effects such as in the XOR problem. Thus, as part of future research, different search strategies will be tested for overcoming this disadvantage. One such approach could be the use of nature inspired search strategies. Furthermore, work will be done in the expansion of the algorithm for being directly used on continuous features.

9. Funding and declaration of interests: Conflict of interest: none declared.

10 References

[1] Antoniades A., Georgousopoulos C., Forgo N., *ET AL.*: 'Linked2safety: a secure linked data medical information space for semantically-interconnecting EHRs advancing patients' safety in medical research'. IEEE 12th Int. Conf. on Bioinformatics and Bioengineering (BIBE), Larnaka, Cyprus, November 2012, pp. 517–522

[2] Fung B.C.M., Wang K., Chen R., *ET AL.*: 'Privacy-preserving data publishing: a survey of recent developments', *ACM Comput. Surv.*, 2010, **42**, pp. 14:1–14:53

[3] Samarati P., Sweeney L.: 'Generalizing data to provide anonymity when disclosing information'. Principles of Database Systems (PODS), 1998, vol. **98**, p. 188

[4] Samarati P., Sweeney L.: 'Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression', Tech. rep., Technical report, SRI International, 1998

[5] Xu Y., Ma T., Tang M., *ET AL.*: 'A survey of privacy preserving data publishing using generalization and suppression', *Appl. Math*, 2014, **8**, (3), pp. 1103–1116

[6] Sweeney L.: 'Guaranteeing anonymity when sharing medical data, the Datafly system'. Proc. of the AMIA Annual Fall Symp., American Medical Informatics Association, 1997, p. 51

[7] LeFevre K., DeWitt D.J., Ramakrishnan R.: 'Incognito: Efficient full-domain *k*-anonymity'. Proc. of the 2005 ACM SIGMOD Int. Conf. on Management of data, ACM, 2005, pp. 49–60

[8] Fung B., Wang K., Yu P.S.: 'Top-down specialization for information and privacy preservation'. Proc. 21st Int. Conf. on Data Engineering, 2005, ICDE 2005, 2005, pp. 205–216

[9] Fung B., Wang K., Wang L., *ET AL.*: 'A framework for privacy-preserving cluster analysis'. IEEE Int. Conf. on Intelligence and Security Informatics, 2008, ISI 2008, 2008, pp. 46–51

[10] Fung B., Wang K., Yu P.S.: 'Anonymizing classification data for privacy preservation', *IEEE Trans. Knowl. Data Eng.*, 2007, **19**, (5), pp. 711–725

[11] Aggarwal C.C.: 'On *k*-anonymity and the curse of dimensionality'. Proc. of the 31st Int. Conf. on Very Large Data Bases, VLDB Endowment, 2005, pp. 901–909

doi: 10.1049/htl.2015.0050

[12] Sweeney L.: 'Achieving *k*-anonymity privacy protection using generalization and suppression', *Int. J. Uncertain. Fuzziness and Knowl.-Based Syst.*, 2002, **10**, (05), pp. 571–588
[13] Zhong S., Yang Z., Chen T.: '*k*-anonymous data collection', *Inf. Sci.*, 2009, **179**, (17), pp. 2948–2963
[14] Kisilevich S., Rokach L., Elovici Y., *ET AL.*: 'Efficient multidimensional suppression for *k*-anonymity', *IEEE Trans. Knowl. Data Eng.*, 2010, **22**, (3), pp. 334–347
[15] Matatov N., Rokach L., Maimon O.: 'Privacy-preserving data mining: a feature set partitioning approach', *Inf. Sci.*, 2010, **180**, (14), pp. 2696–2720
[16] Lichman M.: 'UCI machine learning repository', 2013
[17] Fayyad U.M., Irani K.B.: 'Multi-interval discretization of continuous-valued attributes for classification learning'. Int. Joint Conf. on Artificial Intelligence (IJCAI), 1993, pp. 1022–1029
[18] DecenciÃ¨re E., Zhang X., Cazuguel G., *ET AL.*: 'Feedback on a publicly distributed database: the Messidor database', *Image Anal. Stereol.* , 2014, **33**, pp. 231–234
[19] Kannas C., Achilleos K., Antoniou Z., *ET AL.*: 'A workflow system for virtual screening in cancer chemoprevention'. IEEE 12th Int. Conf. on Bioinformatics & Bioengineering (BIBE), 2012, pp. 439–446
[20] Cortes C., Vapnik V.: 'Support-vector networks', *Mach. Learn.*, 1995, **20**, (3), pp. 273–297
[21] Aha D.W., Kibler D., Albert M.K.: 'Instance-based learning algorithms', *Mach. Learn.*, 1991, **6**, (1), pp. 37–66
[22] Quinlan J.R., C4. 5: Programs for Machine Learning, Morgan Kaufmann, 1993, vol. 1
[23] Breiman L.: 'Random forests', *Mach. Learn.*, 2001, **45**, pp. 5–32
[24] Hornik K., Buchta C., Zeileis A.: 'Open-source machine learning: R meets Weka', *Comput. Stat.*, 2009, **24**, (2), pp. 225–232
[25] Liaw A., Wiener M.: 'Classification and regression by randomforest', *R News*, 2002, **2**, (3), pp. 18–22
[26] Antoniades A., Keane J., Aristodimou A., *ET AL.*: 'The effects of applying cell-suppression and perturbation to aggregated genetic data'. IEEE 12th Int. Conf. on Bioinformatics & Bioengineering (BIBE), 2012, 2012, pp. 644–649