TRANSACTIONS ON DATA PRIVACY 10 (2017) 145 –174

A Multi-dimensional Privacy-aware Evaluation Function in Automatic Feature Selection

Yasser Jafer[1], Stan Matwin[23], Marina Sokolova[124]

[1] School of Electrical Engineering and Computer Science, University of Ottawa, Canada

[2] Institute for Big Data Analytics, Dalhousie University, Canada

[3] Institute of Computer Science, Polish Academy of Sciences, Poland

[4] Faculty of Medicine, University of Ottawa, Canada

E-mail: {yjafer, sokolova}@uottawa.ca, stan@cs.dal.ca

Received 15 September 2015; received in revised form 20 January 2017 and 4 April 2017; accepted 4 October 2017

Abstract. Feature selection is based on the notion that redundant and/or irrelevant variables bring no additional information about the data classes and can be considered noise for the predictor. As a result, the total feature set of a dataset could be minimized to only a few features containing maximum discrimination information about the class. Classification accuracy is used as the evaluation measure in guiding the feature selection process. At the same time, such measure does not take into account the privacy of the resulting dataset. In this work, we introduce *E*(S) a multi-dimensional privacy-aware evaluation function in automatic feature selection that enables the DH to select and eventually release the best subset according to its desired efficacy (e.g., accuracy), privacy, and dimensionality of the resulting dataset.

Keywords. Feature selection, privacy, data mining, classification, evaluation measure

1 Introduction

Datasets, including financial, medical, revenue records, and so on, create new opportunities for knowledge discovery. We need to ensure, however, that personally identifiable information is protected in them; yet, the analytical value of the datasets is preserved.

High dimensionality, an essential characteristic of many large data sets, makes learning tasks (e.g., classification) challenging, and therefore dimensionality reduction tools such as feature selection become indispensable. Dimensionality reduction can be achieved by the elimination, extraction, and engineering of features. By reducing the number of features, feature selection aims to enhance the understandability of data, lower the computational cost, reduce a negative impact of the curse of dimensionality, and improve the predictive performance of learning algorithms (Chandrashekar and Sahin, 2014). In this work, we consider that features and attributes are synonymous.

PPDP (Privacy Preserving Data Publishing) is concerned with developing tools and methods that enable the publishing of data in an insecure environment. PPDP aims at publishing modified data such that while individuals' privacy is preserved, the published data remains practically

useful (Fung et al, 2010). In general, PPDP does not make assumptions about how the data will be used. In the past, different privacy-preserving methods have been proposed aiming to sustain the data utility for certain data mining tasks (Fung et al., 2005; Iyengar, 2002; LeFevre et al., 2006; Wang et al., 2004). Similarly, our work addresses the situation when the data is published for the purpose of being explored by data mining methods.

We assume a scenario that consists of a Data Holder (DH) who holds the original data on the one hand and, a Data Recipient (DR) who wants the data in order to apply certain data mining tasks on the other hand. In our work, we assume that the DR wants to classify the dataset. The DH uses automatic feature selection and publishes a customized dataset, which takes the intended analysis task of the DR into consideration. Since the dataset is going to be published in an insecure environment, in practice, it will also be available to an attacker. However, since the dataset is tailored for a given analysis task, if the attacker uses the same dataset to do, say, clustering analysis instead, the results would be misleading and of less or no value. The DH and the DR could be a hospital and a research center respectively. As such, we assume that the DR is the legitimate user of the released dataset.

Generally, the goal of automatic feature selection is to obtain the most optimal feature set that provides for the algorithm's best performance. In our work, we incorporate privacy considerations "during" the feature selection. As such, the feature selection process is guided in a way that the DH can predict the amount of inference of an individual sensitive attribute to the attacker who has access to sophisticated data mining tools. We introduce a multi-dimensional privacy-aware evaluation function in order to evaluate the potential privacy-aware attribute subsets and obtain the best subset according to the DH's desired preferences of efficacy (e.g., accuracy), privacy, and dimensionality of the resulting dataset.

The contribution of this work is twofold: First, we consider a case where the target class and the sensitive attribute are the same, i.e., C = SA. This implies that the goal of the adversary and the goal of the legitimate user/analyzer of the dataset is the same. This is an extension of the methodology proposed in (Jafer et al., 2015a) in which we consider C! = SA. In that work, we assumed that the goal of the adversary and the goal of the legitimate user/analyzer is different. Second, it introduces a new multi-dimensional measure for a privacy aware feature selection. Rather than obtaining only a list of candidate privacy-aware attribute subsets, we introduce an evaluation function $E(S)$ in order to select a single best subset according to the DH's desired accuracy, privacy, and dimensionality (number of attributes in the resulting subset) preferences.

In the remaining parts of this work, we first start by discussing feature selection and privacy in Section 2. We then, introduce the main characteristics of the proposed methodology in (Jafer et al., 2015a) in Sections an 3d We show how this methodology .4could be adapted to the case of C = SA with slight modification (Section 4). We then introduce the multi-dimensional evaluation function in Section We show the experimental results (Section .56) of a two dimensional plot of candidate privacy-aware attribute subsets (corresponding to C!=SA), followed by the results of implementing the proposed $E(S)$ evaluation function.

2 Feature Selection and Privacy

In this section we discuss some background information about feature selection and privacy and related work.

The main goal in feature selection is to obtain an *optimal* feature set which can be a set that consists of all strongly relevant features and weakly relevant features that are non-redundant (Yu and Liu, 2004). Tsamardinos and Aliferis (Tsamardinos and Aliferis, 2003) proposed that the feature selection problem must include a classifier (or an ensemble of classifiers) in addition to a performance metric. As such, the optimal attribute subset is considered the one which would

maximize the performance metric, and yet has the minimum cardinality. Two main categories of automatic feature selection techniques are *wrappers* and *filters*. Wrappers conduct a search for a good attribute subset using the learning algorithm itself as part of the evaluation function. In contrast, filters use general characteristics of the data in order to evaluate attributes. The main advantage of wrappers is that, they usually result in higher performance compared with filters, and in cases where performance is the main objective they become very practical. The main disadvantage of wrappers is that due to their frequent interaction with the predictor, they turn out to be slower.

Most work in the area of PPDP does not make any assumption about an intended analysis task applied on the dataset. In many domains such as healthcare, finance, etc., however, it is possible to identify the analysis task beforehand. In general, considering a particular publishing scenario is a fruitful direction and leads to identifying the best algorithm based on the scenario at hand (Ayala-Rivera et al., 2014). Incorporating such knowledge of the ultimate analysis task may improve the quality of the anonymized data while protecting the privacy of individuals. However, even when the ultimate analysis task is defined, say, building a classification model, one challenging issue is handling high-dimensional data. As the dimensionality of data increases, data analysis such as classification becomes substantially harder. In some cases, data becomes so sparse leading to the "curse of dimensionality" (Powell 2007). It is possible that in the case of classification the available training data may be very small. Therefore, there will be very few data objects in order to create a reliable model which assigns a class to all possible objects. As a result, large numbers of features may lead to lower classification accuracy. Furthermore, a dataset with high dimensionality is considered a serious problem in many classification techniques due to the associated memory usage and computational cost (Janecek et al. 2008).

Therefore some recent works in (Jafer, 2014) and (Jafer et al., 2014b) have considered using feature selection, a well known dimensionality reduction tool, for privacy preserving purposes. In (Jafer et al., 2014b), it was shown that when feature selection is used off-the-shelf (and without any privacy considerations) it may be utilized towards privacy protection. The justification is that, removing irrelevant and redundant attributes implicitly results in the removal of some of the quasi-identifier attributes. In other words, it is very possible that some of the eliminated irrelevant and/or redundant features are quasi-identifiers, so they get removed by default. In order to single out an individual in a dataset, the set of *all* Quasi-Identifier (QI) attributes is needed. When some of these QIs get removed by feature selection, identity disclosure becomes more difficult. Moreover, with fewer QI attributes less modification is required and hence more details of relevant attributes are preserved. This, as shown in (Jafer et al., 2014b), results in higher utility at the same privacy level (i.e., higher utility at same k in case of k-anonymity (Sweeney, 2002)). The work in (Jafer et al., 2014b) considers the role of feature selection as a privacy preserving tool in the context of k-anonymity.

In (Jafer et al., 2014c), feature selection combined with k-anonymity was utilized in order to increase the count of the resulting contingency tables. In the non-interactive model, one way of achieving differential privacy (Dwork, 2006) is to obtain a contingency table and then, add noise to it. The work showed that, when k-anonymity is preceded by feature selection, the resulting contingency tables built from k-anonymized dataset would have higher counts. As such, when Laplace noise is added to these counts to achieve ε-differential privacy, the utility of the published dataset is preserved. Although the cited works consider feature selection as a privacy-preserving tool, the feature selection procedure itself is not privacy-aware.

Recently, privacy-aware filters and wrappers have been addressed in (Jafer et al., 2014a) and (Jafer et al., 2015b) respectively in which filters and wrappers are turned into privacy-aware processes. These works have two common aspects. First, they mainly focus on protecting against identity disclosure with the notion of QI attributes clearly identified and utilized in the solution.

Second, they do not consider the feature selection's very evaluation measure. Rather, a privacy layer is built on top of the already existing automatic feature selection process.

In a recent work (Jafer et al., 2015a), we presented a two-dimensional view of potential privacy-aware subsets of attributes considering the accuracy and the privacy of a given subset simultaneously. In other words, we modify the very evaluation process to make it privacy-aware and do not make any assumption about the QI attributes.

We build our current work upon (Jafer et al., 2015a) and further extend it by enabling the DH to choose a single best subset according to his/her desired trade-off. We adapt (with modification) part of the proposed methodology in (Jafer et al., 2015a) in order to obtain the candidate privacy-aware attribute subsets.

3 Measuring Privacy-preserving Feature Selection

We will set up the stage for introducing the new evaluation function by some background information about the steps that are taken in order to generate privacy-aware candidate attribute subsets. We first start with the case where C!=SA which was extensively discussed in (Jafer et al., 2015a). We then, show how the algorithms introduced in (Jafer et al., 2015a) could be slightly modified in order to adapt the case where C=SA.

3.1 Basic Notations

We assume that *D* is a dataset with tuple set $T=\{t_1, t_2, \ldots, t_n\}$ and attribute set $A=\{a_1, \ldots, a_m\}$. We also assume that *D* is a subset of some larger population *P*. As such, we have a distribution of the attribute values in *D* that represents the distribution of the population as a whole. We refer to the sensitive attribute (e.g., medical condition of patients) as SA $\in A$. We follow the assumption that, the sensitive attribute will be released without any modification. However, the association between individuals and their sensitive attributes should be kept private. Attribute C $\in A$ is referred to as the target class attribute. Baseline attribute subset *BL* corresponds to a subset of attributes in *A* excluding SA and C. That is, *BL* $= A\backslash(C \cup SA)$. In (Jafer et al., 2015a), we assumed that C and SA are different. In other words, we assumed that the goal of the adversary differs from the legitimate user of the dataset (e.g., researcher). In this work, we focus on the case where C=SA. Baseline accuracy subset *BLC* refers to a subset of all attributes except SA (i.e., *A*- SA). Baseline privacy subset *BLP* refers to a subset of all attributes except C (i.e., *A* - C). Figure 1 illustrates these subsets. A dataset *DBLP* is the projection of the tuple set *T* onto the attributes in *BLP* and a dataset *DBLC* is the projection of tuple set *T* onto the attributes in *BLC*. These notations will be used throughout the paper.

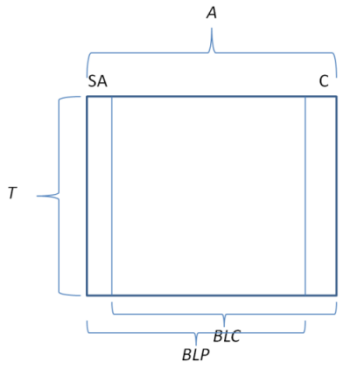

Figure 1: Dataset D and the projection of attributes (case C!=SA)

3.2 Defining the PBI Measure of Privacy

The measure of privacy *Privacy Breach Increase* (*PBI*) (Jafer et al., 2015a) is inspired by the notion of "empirical privacy" proposed in (Cormode et al., 2013). Similar to "empirical privacy", our measure represents the precision with which the attacker can infer the sensitive values of individuals from released data. It is inspired by a widely adopted notion of privacy breach which refers to the correct posterior inference of an adversary about sensitive values in the data. Such measure considers a sophisticated attacker who will use data mining tools for his/her attack. The empirical privacy studies the increase/decrease of attacker's inference ability as a function of the amount of anonymization (Cormode et al., 2013). For example, how much such inference changes when the dataset is *k*=100 anonymized vs. when *k*=1000 anonymized. In the *PBI* measure, privacy breach increase (or decrease) is considered as a function of the selected attributes in the final dataset (or implicitly, the number of eliminated attributes). The work in (Cormode et al., 2013) considers an adversary who would use the output *Anon*(*D*) (where *Anon* refers to an anonymization mechanism) to build a classifier in order to attack anonymized data. With the assumption that feature selection is done prior to anonymization, we consider an attacker who will use only the selected features (compared with the complete feature set i.e., baseline dataset) in order to build a classifier to predict the sensitive attribute. We study the impact of the reduction of selected attributes on the overall ability of the attacker to correctly predict the value of the sensitive attribute.

We assume that *DS* represents the projection of a selected attributes set *S* on the dataset *D* (where $S \subset A \backslash C$). This projected dataset is used to build a model of the data in order to classify SA. In our terminology, $Acc(DS)$ and $Acc(DBLP)$ refer to the accuracy of correctly predicting the value of SA. We build a classifier using *DS* and *DBLP* respectively.

Definition 1: Given dataset *D*, a projected dataset *DS*, and the baseline dataset $DBLP$, we compute *Privacy Breach Increase (PBI)* as

$$PBI(DS) = \left(\frac{Acc(DS)}{Acc(DBLP)}\right) - 1 \quad (1)$$

We utilize *PBI* as a measure used to represent privacy in terms of how easy/difficult it is to correctly predict the sensitive attributes using a given attribute set. Implementing formula (1), the

following outcomes are possible: if *PBI* is positive, this is an indication that the privacy risk of *DS* is higher than the baseline dataset. Likewise, a negative value of PBI indicates that the privacy risk of *DS* is lower than the baseline dataset. PBI is computed as the percentage of the baseline privacy. The minimum privacy breach is referred to as distribution privacy *DistP*. The amount of allowed *PBI* of a given dataset is chosen by DH. For example, if the *PBI* associated with a given attribute subset is (-10%), it means that it becomes (10%) more difficult for the attacker to correctly predict the correct value of SA for a given individual when compared with attacking the baseline dataset. The distribution privacy is simply the distribution of the sensitive attributes which represents the population distribution and is known to the adversary (since the assumption is that the sensitive attributes will be released). If we are able to publish a dataset in which the prediction of the sensitive attribute is not more than the sensitive attribute distribution *DistP*, even if the attacker uses the same dataset in order to infer the values of the sensitive attribute, his/her posterior belief will not change as the result of releasing the dataset. This observation could be related to the notion of ε-differential privacy in which the privacy of any individual should not be substantially (bounded by ε) disclosed as a result of participation in a statistical database.

One remaining question is that, to what extend the proposed method is robust against malicious attacks in which the attacker would explore in-depth understanding of the data (e.g., correlation between attributes, etc.)? Our privacy objective of a selected subset is achieved proportional to the baseline. In other words, we ensure that the privacy breach due to the release of a given dataset (with selected features) is lower than that of the baseline and below a given threshold. To this end, the aim is that to minimize the capability of the attacker in correctly predicting the correct value of SA as a result of dataset release. We insist that, the purpose here is to limit and control the amount of inference of sensitive attribute(s) due to the release of the dataset. Our aim is to control the difference between the prior and posterior belief of the attacker (i.e., before and after seeing the released dataset). If the attacker is able to infer SA using external databases, such inference will be made anyway and regardless of releasing/not releasing the dataset held by DH. In other words, it is DH's responsibility to manage the dataset he/she owns and that needs be released. Furthermore, we assume that the recipient of the dataset (DR) is a well-known and trusted entity for a given analysis purpose.

3.3 Application of PBI

Objective: We want to find attribute subset $S \subset A \backslash (C \cup SA)$ such that (1) the probability of achieving correct posterior inference about the sensitive attribute (i.e., SA) of an individual based on *S* is below a given threshold α (privacy) and (2) the difference between the performance of the baseline attribute subset *BL* and the selected attribute subset *S* (in terms of predicting the target attribute C) is not statistically significant or is in favor of the selected attribute subset (efficacy). Formally, we want to find

$$S|\ (PBI(DS) \leq \alpha) \& \Big(\big(Perf(DS)\ \geq Perf(DBL)\big) ||\ \ (|Perf(DBL) - Perf(DS)|)! = statistically\ significant \Big) \tag{2}$$

From this objective, it is possible to obtain more than one attribute subset which satisfies the above constraints. We refer to these attribute subsets as **candidate privacy-aware attribute subsets**. Depending on the user's preferences, within the space of candidate privacy-aware subsets, four possible regions could be identified which are shown in Figure 2. We want to enable the user to select a trade-off between performance (i.e., classification accuracy) and privacy (i.e., *PBI*) based on his/her requirements and priorities. The user might be willing to give

up some utility in order to gain more privacy and vice versa. Each point in Figure 2 represents a selected attribute subset with two associated values (*Perf*(*DS*), *PBI*(*DS*)).

Perf(*DS*) refers to the accuracy of the attribute subset in predicting the target class C. *PBI*(DS) refers to the increase in the likelihood of privacy breach of the attribute subset S in predicting the sensitive attribute SA compared with the baseline. *DistP_PBI* refers to the *PBI* associated with the distribution of the sensitive attribute in the whole dataset w.r.t the baseline. This is the maximum privacy guarantee and is unaffected by the records in the dataset. These regions are listed as follows:

- **Region I (NW)**: The attribute subsets in this region have a *PBI*(*DS*) and *Perf*(*DS*) higher and lower than the baseline respectively**.** In other words, any other candidate attribute subset in other regions will surpass the attribute subsets in this region because provide either better performance or better privacy or both. NW, in other words, is the worst region to be at**.**
- **Region II (NE)**: The attribute subsets in this region result in higher *PBI*(*DS*). However, they achieve higher performance compared with the baseline.
- **Region III (SW)**: The attribute subsets in this region have performance that is lower than the baseline performance; however, their *PBI*(*DS*) is also lower than the baseline.
- **Region IV (SE)**: The attribute subsets in this region achieve better performance compared with the baseline while incurring less potential privacy breach since their *PBI*(*DS*) is lower than that of the baseline. This is the best region/location to be at *PBI*(*DS*) = *DistP_PBI*.

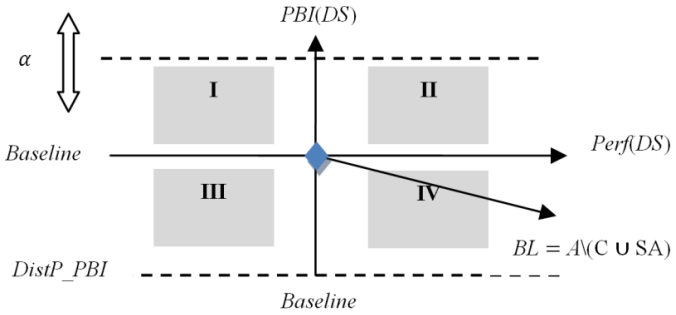

Figure 2: Illustration of performance vs. privacy trade-off.

4 Candidate Privacy-aware Attribute Subset Generating System

The initial privacy-aware feature selection system consists of two subsystems, namely, *Ranker* and *Candidate Subset Generator* (Figure 3). The details of each of these subsystems are provided in the following subsections.

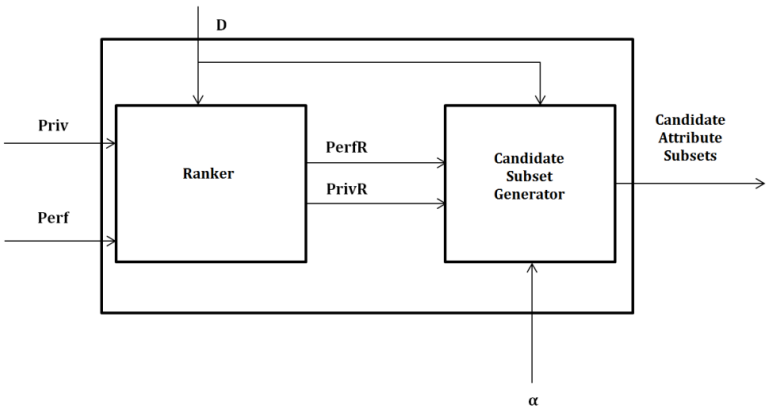

Figure 3: Candidate Privacy-aware Attribute Subset Generating System

4.1 Correlation-aware Attribute Ranking

Each attribute may have a different privacy risk associated with it. We first generate two ranked lists of attributes, namely, a performance rank (*PerfR*) and a privacy rank (*PrivR*). Let us focus on the privacy rank (*PrivR*). We want to determine the impact of each attribute on correctly predicting the sensitive attribute SA. Our assumption is that SA differs from the target attribute C. First, we build a classifier using attributes in $BL \cup SA$ with respect to SA, i.e., to obtain *BLP* (See Section 3.1). In order to find the privacy risk of each attribute, only one attribute is removed at a time. Our algorithm then, finds the projection of tuple set *T* onto the attributes in the remaining attribute set and then calculates the accuracy of the projected dataset with respect to SA. The difference between the accuracies of *DBLP* and the projected dataset is calculated. This difference indicates the impact of removing a given attribute on increasing/decreasing the accuracy of predicting the SA attribute. In the following step, we rank the attributes from lower to higher. When the attribute's ranking is higher it would have a higher contribution in correctly classifying SA, i.e., the attribute is more risky. As such, *PrivR* will consist of attributes that are ranked from lower to higher privacy risk.

We follow the same logic in order to obtain *PerfR*. There is, however, a difference. We assume that the baseline attribute set is *BLA*. We take a step by step approach. We start from the complete list of attributes. Attributes are then, eliminated one-by-one. After that, we build a classifier using the remaining attributes in order to predict the target class attribute C. Finally, the difference between the accuracies of *DBLA* and the projected dataset is calculated and the ranked list of attributes according to their performance attribution is obtained. *PerfR* ranks the attributes from higher to lower. This algorithm is shown in Figure 4. The output of this algorithm consists of two ranked lists of attributes *PrivR* and *PerfR* which will be used as input to the "Privacy-aware Candidate Subset Generator" algorithm (Section 4.3).

The reason for ranking privacy and performance lists differently (lower to higher, higher to lower respectively) is that when the ranked list is used in the next algorithm, in backward elimination, starting from the last attribute in the ranked list, we tend to, first, remove attributes that have higher privacy risk and lower impact on performance.

4.2 Searching for Candidate Attribute Subsets

In many search problems, the path to the goal is irrelevant (Russell et al., 2010). As such, it is possible to consider different types of algorithms that do not give importance to the path at all. In

other words, what is important is the solution state. Local search algorithms operate using a single node and move only to neighbors of that node. There are two main advantages associated with the local search algorithms as stated in (Russell et al., 2010): (1) they need very little memory because the paths followed are not retained. (2) They usually find reasonable solutions in large or infinite state spaces where it is not suitable to use systematic algorithms. In addition to finding goals, one of the main benefit(s) of using local search algorithms is that they are very useful in order to solve optimization problems. In optimization problems, the goal is to find the best state according to given objective function.

Input: Dataset D (A attributes and T tuples, sensitive attribute SA, target attribute C)

$BL = A \setminus (SA \cup C)$ $BL = \{BL_1, BL_2, \ldots, BL_k\}$
$BLP = BL \cup SA$
$BLC = BL \cup C$
$DBLP = proj(D, BLP)$
$DBLC = proj(D, BLC)$
$Perf\_BLP = Acc(DBLP)$
$Perf\_BLC = Acc(DBLC)$

//This function returns a list of all attributes ranked according to the required order.
Rank((attribute, value), order); order{high_to_low,low_to_high}:

for (i = 1, i ≤ k, i++)
{
$BL_{temp\_i} = BLP - \{BL_i\}$
$DBL_{temp\_i} = proj(D, BL_{temp\_i})$
$BLP_i = Perf\_BLP - Acc(DBL_{temp\_i})$
$PrivR \leftarrow (\{BLi\}, BLP_i)$
}

$PrivR$ = Rank($PrivR$, low_to_high)
for (j = 1, j ≤ k, j++)
{
$BL_{temp\_j} = BLC - \{BL_j\}$
$DBL_{temp\_j} = proj(D, BL_{temp\_j})$
$BLC_j = Perf\_BLC - Acc(DBL_{temp\_j})$
$PerfR \leftarrow (\{BLj\}, BLC_j)$

}
$PerfR$ = Rank($PerfR$, high_to_low)

Output: privacy-based ranked attributes (PrivR), performance-based ranked attributes (PerfR)

Figure 4: Privacy-based and Performance-based Ranking Algorithm.

Objective function is also considered a heuristic cost function. In addition to the local hill-climbing search, there are other local search algorithms such as simulated annealing, local beam

search, and genetic algorithms, which are beyond the scope of this paper. To this end, in the context of search, we identify our search space to be the space of all subsets. Obviously, with n attributes such search will include 2^n possibilities. After all, this is the main reason why wrappers usually follow heuristic approaches (e.g., best-first with forward selection or backward elimination). Our search for a privacy-aware candidate subset is informed or heuristic in the sense that the subsets are selected based on an evaluation function. This evaluation function is guided by the rank of attributes in the *PrivR* and *PerfR* list since the order in which the attributes are eliminated is dictated by these ranks. In our search we aim to remove one attribute at a time and therefore we conduct a local search. Therefore, we essentially follow a stepwise backward elimination procedure. We use backward elimination in order to preserve features whose usefulness requires other features. The very definition of stepwise backward elimination is that we start with the full set of attributes and then, at each step, remove the worst attribute that remains in the set. Such selection of the "worst" attribute is dictated by the *PerfR* and *PrivR*. Our search is guided by both *PerfR* and *PrivR*. In the case of *PerfR* we rank the attributes based on their relevance with respect to the target class. Therefore, worst attributes refer to the ones that are the least predictive. When the search is guided by the *PrivR* list, worst attributes refer to the ones that are most predictive of the sensitive attribute. After each removal the remaining subset is tested against the accuracy and the privacy requirements. If both requirements are met the given subset is added to the list of candidate subsets. If not, the last removal is cancelled and the next attribute in the ranked list is removed.

In our search methodology, similar to hill climbing, only one node is expanded at a time with the goal of finding local maxima. This is different than best first search where the goal is to find global maxima and either to find an optimal solution or to empty the complete open list. Such inability to find an optimal solution is well understood in the context of privacy-aware feature selection. The very definition of optimal feature subset in the case of regular wrapper is to obtain a feature subset with minimum number of attributes that has the highest predictive accuracy. From formula (2), in a privacy-aware feature selection, we conduct a search for subsets that satisfy our efficacy and privacy requirements simultaneously. In most of the cases, however, it is not possible to find a given subset with the highest efficacy and lowest *PBI* simultaneously. This objective becomes less realistic if we add a third requirement of having minimum number of attributes. Therefore, our search aims to obtain a list of subsets (hence, candidate subsets) that satisfy the efficacy and privacy requirements.

Obtaining a single subset is an extra step beyond generating candidate privacy-aware attribute subsets by means of incorporating accuracy, privacy, and dimensionality weights based on the DH's preferences. As such, it becomes possible to narrow down the search for a feature subset to a single subset according to the DH's preferences. Introducing a multi-dimensional evaluation function is the main contribution of this work. We start with the background information that precedes the definition of the proposed multi-dimensional evaluation function.

To the best of our knowledge, no other work has considered a privacy-aware evaluation function in feature selection. In the existing solutions, feature selection is performed separately and hence, no comparison study was conducted. Furthermore, in the proposed privacy-aware feature selection measure, the process of feature selection is tied to a search for local maxima and not global maxima. The reason is that, as was mentioned earlier in this section, the search is guided by a privacy factor (via *PrivR*) and a utility factor (via *PerfR*) and we are not searching for the best subset, rather several candidate subsets that satisfy utility and privacy thresholds. We then use the evaluation function in order to find the best subset according to some preferences. Therefore, execution time is implicitly improved when the technique is compared with wrapper-based feature selection where searching the entire space for the best subset turns out to be very slow, especially in large datasets.

4.3 Candidate Attribute Subsets Generator

Following the step of finding ranked lists of attributes in *PerfR* and *PrivR*, we implement a backward elimination search technique which is guided by these ranked lists. The main reason for calculating two ranked lists of *PerfR* and *PrivR*, as mentioned in the previous section, is to identify subsets of attributes that consider either performance or privacy as their main priority. With such an approach, we provide the DH/publisher with greater flexibility when selecting the final subset of attributes.

The input to our candidate attribute subset generator algorithm includes dataset *D*, the *PrivR* list, the *PerfR* list, the value of *Perf_BLC* (Performance of baseline attribute set *BL* with respect to the target class C), and the value of *Perf_BLP* (Performance of baseline attribute set *BL* with respect to SA). We start from the bottom of a given list. This list could be either *PrivR* or *PerfR*. We then, eliminate the attributes, attribute by attribute. With each elimination, both privacy and performance constraints are checked. If both constraints are satisfied this attribute is removed and the search proceeds to the next one. If one of the constraints is violated, we do not eliminate this attribute and go to the next attribute in the list. This process is repeated until any further elimination would violate one of the constraints or when the attribute list is used up.

Using the output of privacy-aware candidate attribute subset generator algorithm we populate the PBI(DS) vs. Perf(DS) diagram. Therefore, each point in that diagram (Figure 6) represents a subset of attributes. It was mentioned earlier in this section that our favorite subsets of attributes are in region IV. Presumably, we are interested in a candidate attribute subset in IV with the minimum number of features. DH is responsible for controlling the amount of privacy prior to release of a dataset (with selected feature vector). As such, the DH needs to ensure that there is no other subset that is better (in terms of PBI) that the ones selected in region IV.

We selected the *Pima* dataset in order to show our approach in detail (Jafer et al., 2015a). We consider that the goal is to build a C4.5 classifier. The attacker's goal is to build a classifier in order to infer the sensitive attributes of the individuals. The *Pima* dataset consists of 768 records and nine attributes, namely, *Preg*, *Plas*, *Pres*, *Skin*, *Insu*, *Mass*, *Pedi*, *Age*, and *Class* attribute. These attributes are defined in the UCI repository. The attributes refer to number of times pregnant, plasma glucose concentration at 2 hours in an oral glucose tolerance test, diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight in kg/(height in m)^2), Diabetes pedigree function, Age (years), and Class variable (0 or 1) respectively. We assume that *Preg* is the SA, in other words, the number of times pregnant is consider sensitive attribute. The class attribute refers to whether a given patient has/does not have diabetes.

Input: Dataset *D* (*A* attributes and *T* tuples, sensitive attribute SA, target attribute C), {*PrivR*|*PerfR*}, *Perf_BLC* (Performance of baseline with respect to C), *Perf_BLP* (Performance
of baseline with respect to SA)

RankedList= {R_1, R_2, …, R_k} // depending on the input either {*PrivR*|*PerfR*}
n= 0;
for (i = k, i>0, i--)
{
S = *RankedList* – {R_i}
S_C = $S \cup C$
S_SA = $S \cup SA$

```
        DBS_C = proj(D, S_C)
        DBS_SA= proj(D, S_SA)
        if ((Acc(DBS_C) ≥ Perf_BLC        ||
(|Perf(DBS_C) − Perf_BLC|)! = statistically significant)
        {
           PBI(DBS_SA) = ((1- Err(DBS_SA))/ Perf_BLP) -1
           if(PBI(DBS_SA) ≤ α ) {
              Candidate_Privacy_Aware_Subset[n] ← (S, Acc(DBS_C ), PBI(DBS_SA))
               n++
           }
        } else { // the elimination of {Ri} violates the utility requirements, we need
                 // to put it back and continue with the next attribute in the ranked list
           RankedList = RankedList ∪ {Ri}
        }
  }
Output: Candidate_Privacy_Aware_Subset
```

Figure 5: Privacy-aware Candidate Subset Generator.

The first step includes obtaining *PerfR* and *PrivR* ranks. These ranks are as follows:

PerfR = {Plas,Mass,Pres,Skin,Age,Pedi,Insu}
PrivR = {Skin,Insu,Pres,Mass,Pedi,Plas,Age}

After obtaining the ranks, following backward elimination strategy guided by these ranks, we eliminate the attributes step-wise and obtain the associated Perf(DS) and PBI(DS) for the remaining attribute subset. These results corresponding to *PerfR* and *PrivR* are shown in Table 1 and Table 2 respectively.

Following a feature's elimination we run a t-test and record the difference in the performance of the projected dataset from the remaining attribute subset and the baseline. Hereafter, when we refer to the PBI(DS) and Perf(DS) of attribute subset we implicitly refer to the projected dataset where its feature vector includes the subset S. The baseline performance for this dataset is 73.70%. If the performance is higher than the baseline or does not differ significantly, this is an indication that this attribute subset satisfies our performance requirements. In other words, the combination of the attribute subsets in Table 1 and Table 2 refers to the list of attribute subsets that yield acceptable performances. ⊕/⊖ refers to a significantly higher/lower performance compared to the baseline performance.

Table 1: Pima - C4.5. PerfR-based candidate attribute subsets.

Subset S	**Perf(DS)**	***p*-value**	**PBI(DS)**
Plas_Mass_Pres_Skin_Age_Pedi (A)	75.26⊕	0.044	0.97
Plas_Mass_Pres_Skin_Age (B)	75.52⊕	0.009	1.29
Plas_Mass_Pres_Skin (C)	73.83	0.931	-3.08
Plas_Mass_Pres (D)	74.35	0.599	-3.08
Plas_Mass (E)	74.48	0.525	-2.75
Plas (F)	73.04	0.626	-2.75

Note that in Table 2, elimination of the Plas attribute significantly reduces the performance compared with the baseline. In such a case, following the algorithm, Plas is put back and the next attribute in the PrivR rank i.e., Pedi is eliminated and candidate attribute subset *H* is obtained.

Table 2: Pima - C4.5. PrivR-based candidate attribute subsets.

Subset S	**Perf(DS)**	***p*-value**	**PBI(DS)**
Skin_Insu_Pres_Mass_Pedi_Plas (G)	73.96	0.8544	-3.24
~~Skin_Insu_Pres_Mass_Pedi~~	~~66.15~~	~~0.0015~~	N/A
Skin_Insu_Pres_Mass_Plas (H)	73.57	0.9318	-3.08
Skin_Insu_Pres_Plas (I)	72.91	0.6115	-3.08
Skin_Insu_Plas (J)	73.30	0.7542	-3.08
Skin_Plas (K)	73.30	0.7542	-3.08

The corresponding two-dimensional diagram of the candidate attribute subsets is shown in Figure 6. Each point in this diagram refers to an attribute subset. For consistency and simplicity an alphabetic letter is given to each attribute subset. It is possible to identify the four regions discussed earlier in Figure 2. The dotted line in Figure 6 refers to α. We assume that α is equal to the majority class distribution of SA i.e., *DistP_PBI*. For abbreviations, each of the subsets is identified with a corresponding alphabetic letter. Technically, the most ideal attribute subsets include *H*, *C*, *G*, *D*, and *E*. We are interested in the attribute subset with fewer numbers of attributes. The number of attributes associated with *H*, *C*, *G*, *D*, and *E* are 5, 4, 6, 3, and 2 respectively. Therefore, we may select *E* = {Plas, Mass} as the final attribute subset.

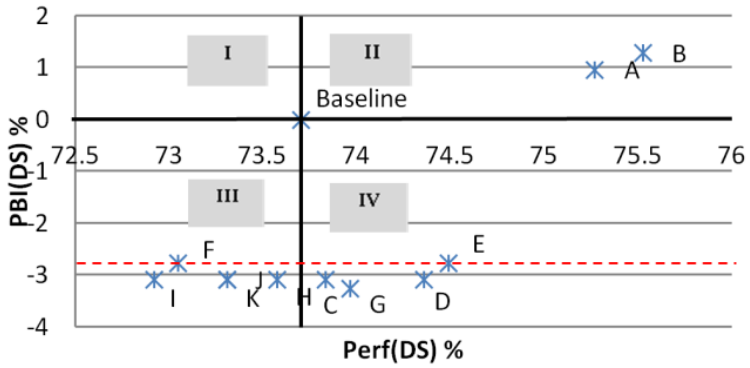

Figure 6: Pima dataset - C4.5 - PBI(DS) vs. Perf(DS).

For comparison purposes, let us consider a generic wrapper feature selection that is not privacy-aware. The selected features using best-first greedy search (with both forward and backward elimination directions) includes {Plas, Pres, Mass, Age}. The performance of the projected dataset using these features is 75.78% which is higher than that of E (i.e., 74.48 %). However, the PBI associated of WFS (Wrapper-based Feature Selection) is 2.43% which is beyond our accepted α. In fact, when generic wrapper is used, the DH cannot have any control over the PBI of the resulting dataset. However, with the privacy-aware evaluation measure the amount of privacy breach increase can be controlled and managed by DH.

4.4 Extension of PBI and Ranker for the case where (C=SA)

So far, we assumed that the goal of the attacker and the legitimate user of the dataset is different (hence, C!=SA). It is possible that these goals are the same in which both try to predict the value of C, one for legitimate purposes and the other for malicious purposes. The algorithms discussed so far could be modified slightly to adapt to such change. For simplicity, we assume that our dataset format is represented in Figure 7.

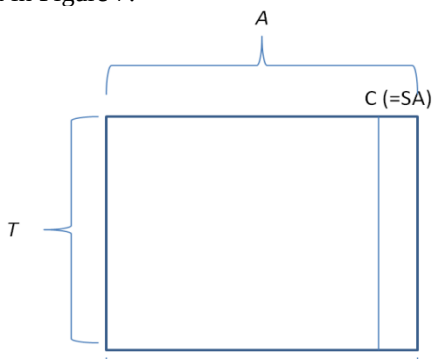

Figure 7: Dataset D and the projection of S attributes (case C=SA).

As before, the assumption is that the sensitive attribute will be released without any modification. However, the association between individuals and their sensitive attributes should be kept secret. Attribute C $\in A$ represents the target class attribute. Baseline attribute subset *BL* refers to a subset of attributes in *A* excluding SA and C. That is, *BL* $=$ *A*\C.

Our modified measure of privacy i.e., *Privacy Breach Increase* is a slight alternation of the *PBI* measure introduced in (Jafer et al., 2015a) for the case where C = SA.

Let *DS* represent the projection of a selected attributes set *S* on the dataset *D* (where *S* $\subset$ *A*\C). We use *DS* to build a model of the data in order to classify C. Assume that Acc(*DS*) and Acc(*DBL*) refer to the accuracy of correctly predicting the value of SA when the classifier is built using *S* and *BL* respectively.

Definition 2: Given dataset *D*, a projected dataset *DS*, and the baseline dataset DBL, we compute the modified *PBI* as

$$\text{Modified } PBI(DS) = \left(\frac{Acc(DS)}{Acc(DBL)}\right) - 1 \tag{3}$$

Furthermore, contrary to the case of C!=SA (in which *PrivR* and *PerfR* are independent), when C=SA, *PerfR* and *PrivR* will be dependent and will have a reverse relation. The methodology of generating privacy-aware candidate attribute subsets in both cases is the same. The main idea being that a two-dimensional plot of Perf(DS) vs. PBI(DS) is obtained and the four regions explained are populated as discussed in Section 3.3.

5 Towards A Multi-dimensional Privacy-aware Evaluation Function

The *Evaluator* Subsystem (which is an extension of the subset shown in Figure 3) implements the proposed multi-dimensional privacy-aware evaluation function *E*(S). This extension is shown in Figure 8. We consider three factors, namely, the performance associated with a given subset *S*, i.e., *Perf*(S), the privacy associated with *S*, i.e., *Priv*(S), and the number of attributes in *S*, i.e., *Num*(*S*).

Ideally, within the list of candidate subsets, we are looking for a subset which has the best *Perf*(S) (i.e., **argmax**(Perf(S))), the best privacy (in the case of *PBI* as the privacy measure, the lowest *PBI*) (i.e., **argmin**(PBI(S))), and the least number of attributes (i.e., **argmin**(Num(S))).

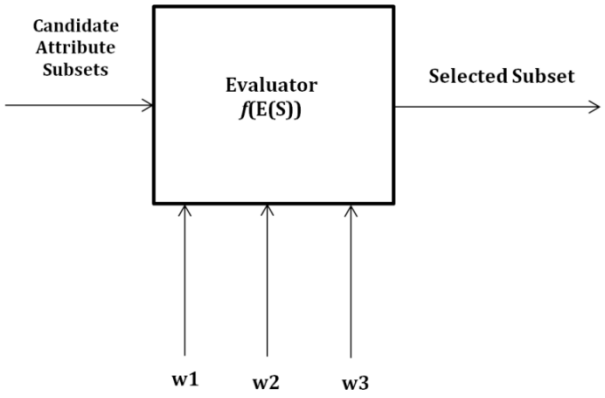

Figure 8: The Evaluator subsystem.

In practice, a given subset might satisfy none, some, or all of these combinations. We can list the possible combination of these factors being either true or false as follows. In binary encoding, (2^3 =) eight combinations may exist. These combinations are shown in Table 3.

It is very likely that we obtain three datasets each satisfying only one condition. In fact, the majority of cases belong to case 'A'. Let us use a running example (the Pima dataset from the UCI repository and the case of C4.5 classifier) in order to discuss this observation. After obtaining a list of candidate subsets of attributes, the number of attributes *Num*, the privacy breach increase *PBI*, and the performance *Perf* associated with each of the candidate subsets are recorded. The results are shown in Table 4. We keep a baseline dataset for reference and comparison purposes.

Table 3: 2^3 combinations of the three identified factors.

Combination	argmax(Perf(S))	argmin(PBI(S))	argmin(Num(S))
A	**F**	**F**	**F**
B	**F**	**F**	**T**
C	**F**	**T**	**F**
D	**F**	**T**	**T**
E	**T**	**F**	**F**
F	**T**	**F**	**T**
G	**T**	**T**	**F**
H	**T**	**T**	**T**

Table 4: The Num, PBI, and Perf associated with candidate subsets (Pima-C4.5). The corresponding combination category is shown in the last column.

Subset	Num	PBI	Perf	Combination
baseline	**8**	**0**	**73.835**	**A**
plas_mass_preg_skin_pedi_age_pres	**7**	**1.409**	**74.875**	**E**
plas_mass_preg_skin_pedi_age	**6**	**1.231**	**74.744**	**A**
plas_mass_preg_skin_pedi	**5**	**0.889**	**74.491**	**A**
plas_mass_preg_skin	**4**	**0.711**	**74.359**	**A**
plas_mass_preg	**3**	**0.711**	**74.359**	**A**
plas_mass	**2**	**0.882**	**74.486**	**A**
plas	**1**	**-1.072**	**73.043**	**B**
insu_pres_age_pedi_skin_preg_plas	**7**	**-1.942**	**72.401**	**C**
insu_pres_age_pedi_skin_plas	**6**	**-0.708**	**73.312**	**A**
insu_pres_age_pedi_plas	**5**	**-1.238**	**72.920**	**A**
insu_pres_age_plas	**4**	**-1.241**	**72.919**	**A**
insu_pres_plas	**3**	**-0.893**	**73.175**	**A**
insu_plas	**2**	**-1.072**	**73.043**	**A**

We highlight the cells with **argmax**(*Perf*(*S*)), **argmin**(*PBI*(*S*)), and **argmin**(*Num*(*S*)). Therefore, we identify three subsets that each satisfy only one ideal requirement i.e., having the best performance, the lowest PBI, or the least number of attributes. These datasets belong to combinations 'E', 'C', and 'B' respectively. However, the remaining datasets belong to combination 'A' where none of the requirements are satisfied.

Our proposed evaluation function *E*(S) relaxes such restrictions and provides the DH with the flexibility of choosing a given factor (being performance or privacy or dimensionality of the data) over other(s). Since the main goal of feature selection is to reduce the dimensionality of data, we consider the number of attributes in the selected subset to be another factor. Our evaluation function is independent of a specific performance measure or privacy measure. It is a function that combines/blends the two measures in such a way that the DH could exercise its privacy/utility preferences. The proposed measure is called *E*(S) and is obtained as,

$$E(S) = w1.prf + w2.prv + w3.num \quad (4)$$

where *prf*, *prv*, and *num* are the rank of a given subset within the set of candidate subsets with respect to *Perf*(*S*), *PBI*(*S*), and *Num*(*S*) respectively. In other words, rather than considering only the maximum or the minimum value of performance, privacy, or number of attributes of a given subset (which is highly limited as it was discussed earlier), we consider *prf*, *prv*, and *num* associated with each candidate subset within the set of candidate subsets.

*w*1, *w*2, and *w*3 refer to the weights given to each of these factors (ranks). We assume that, $\sum w = 1$, i.e., $w1 + w2 + w3 = 1$. Associating weights with the aforementioned factors has two benefits: First, it makes the combinatory evaluation function generalizable. In other words, we can ignore any of the factors by setting its corresponding weight to 0. Second, it allows the DH to evaluate/select the subsets based on his/her preferences. For example, by setting $w2 = 4 * w1$, the privacy of the attribute subset is given four times the level of importance compared with the performance of the subset, and so on. When no preference is given to any of the factors, the default case refers to $w1 = w2 = w3 = 1/3$. In selecting the values of w1, w2, and w3, we considered different cases based on realistic scenarios. In some cases we are more interested in protecting the privacy to a achieve a given threshold, say *k* value (in *k*-anonymity privacy model). In other cases we want to give higher importance to utility while privacy is preserved at an acceptable level, etc. This is reflected in the experimental results and their corresponding figures/plots as it follows.

Technically, the goal of the *Evaluator* subsystem is to apply the evaluation function and to obtain a single selected subset with the best *E*(*S*) given the selected weights:

$$f(E(S)) = \textbf{argmax}(E(S)) \mid w1, w2, w3) \quad (5)$$

Table 5: The candidate attribute subsets, their corresponding Perf(S), PBI(S), and Num(S) and their ranks (pima-C4.5). prf with respect to performance (lowest being the worst and highest being the best), prv with respect to PBI (highest being the worst and lowest being the best).

Subset	n	Num. (R)	PBI	PBI (R)	Perf	Perf (R)
baseline	**8**	**1**	**0**	**7**	**73.835**	**8**
plas_mass_preg_skin_pedi_age_pres	**7**	**2**	**1.409**	**1**	**74.875**	**14**
plas_mass_preg_skin_pedi_age	**6**	**4**	**1.231**	**2**	**74.744**	**13**
plas_mass_preg_skin_pedi	**5**	**6**	**0.889**	**3**	**74.491**	**12**
plas_mass_preg_skin	**4**	**8**	**0.711**	**5**	**74.359**	**9**
plas_mass_preg	**3**	**10**	**0.711**	**5**	**74.359**	**9**
plas_mass	**2**	**12**	**0.882**	**4**	**74.486**	**11**
plas	**1**	**14**	**-1.072**	**10**	**73.043**	**4**
insu_pres_age_pedi_skin_preg_plas	**7**	**2**	**-1.942**	**14**	**72.401**	**1**
insu_pres_age_pedi_skin_plas	**6**	**4**	**-0.708**	**8**	**73.312**	**7**
insu_pres_age_pedi_plas	**5**	**6**	**-1.238**	**12**	**72.920**	**3**
insu_pres_age_plas	**4**	**8**	**-1.241**	**13**	**72.919**	**2**
insu_pres_plas	**3**	**10**	**-0.893**	**9**	**73.175**	**6**
insu_plas	**2**	**12**	**-1.072**	**10**	**73.043**	**4**

Continuing with our running example (the Pima dataset – C4.5), for each of the candidate subsets (and the baseline dataset) we obtain the *prf*, *prv*, and *num* ranks (Table 5). To examine our evaluation function $E(S)$, we consider three cases. In each of these cases, we want to use $E(S)$ in order to obtain a single subset according to our preferences (applied via weights). In case 1 we assume that ($w1 = w2 = w3$). In case 2, we give performance twice the importance of privacy ($w1 = 2 * w2$), and in case 3, we give privacy three times the importance of performance ($w2 = 3 * w1$) . The $E(S)$ results associated with the candidate attribute subsets is shown in Table 6. We distinguish between two categories of subsets, those obtained based on *PerfR* (shown in the top bracket in Table 6) and those obtained based on *PrivR* (shown in the bottom bracket in Table 6).

From Table 6 we notice that in general when performance is given higher importance compared with privacy, the subsets corresponding to *PerfR* result in higher $E(S)$. On the other hand, when privacy is given more importance compared with performance, subsets corresponding to *PrivR* have higher $E(S)$. This is in line with the observation that, the subsets obtained based on *PerfR* ranking have higher accuracy compared with those which are based on *PrivR* ranking. On the other hand, the *PBI* of the corresponding subsets based on *PrivR* ranking is lower than the corresponding subsets that are obtained based on *PerfR* ranking.

We plot the diagrams associated with subset/$E(S)$ combinations and show the results in Figure 9.

Table 6: The E(S) results associated with candidate subsets corresponding to the selected weights (Pima dataset - C4.5).

	Subset	E(S) (w1 = w2 = w3)	E(S) (w1 = 2w2)	E(S) (w2=3w1)
	baseline	**5.328**	**6**	**6**
PerfR	**plas_mass_preg_skin_pedi_age_pres**	**5.661**	**7.75**	**3.8**
	plas_mass_preg_skin_pedi_age	**6.327**	**8**	**4.6**
	plas_mass_preg_skin_pedi	**6.993**	**8.25**	**5.4**
	plas_mass_preg_skin	**7.326**	**7.75**	**6.4**
	plas_mass_preg	**7.992**	**8.25**	**6.8**
	plas_mass	**8.991**	**9.5**	**7**
	plas	**9.324**	**8**	**9.6**

PrivR

insu_pres_age_pedi_skin_preg_plas	**5.661**	**4.5**	**9**
insu_pres_age_pedi_skin_plas	**6.327**	**6.5**	**7**
insu_pres_age_pedi_plas	**6.993**	**6**	**9**
insu_pres_age_plas	**7.659**	**6.25**	**9.8**
insu_pres_plas	**8.325**	**7.75**	**8.6**
insu_plas	**8.658**	**7.5**	**9.2**

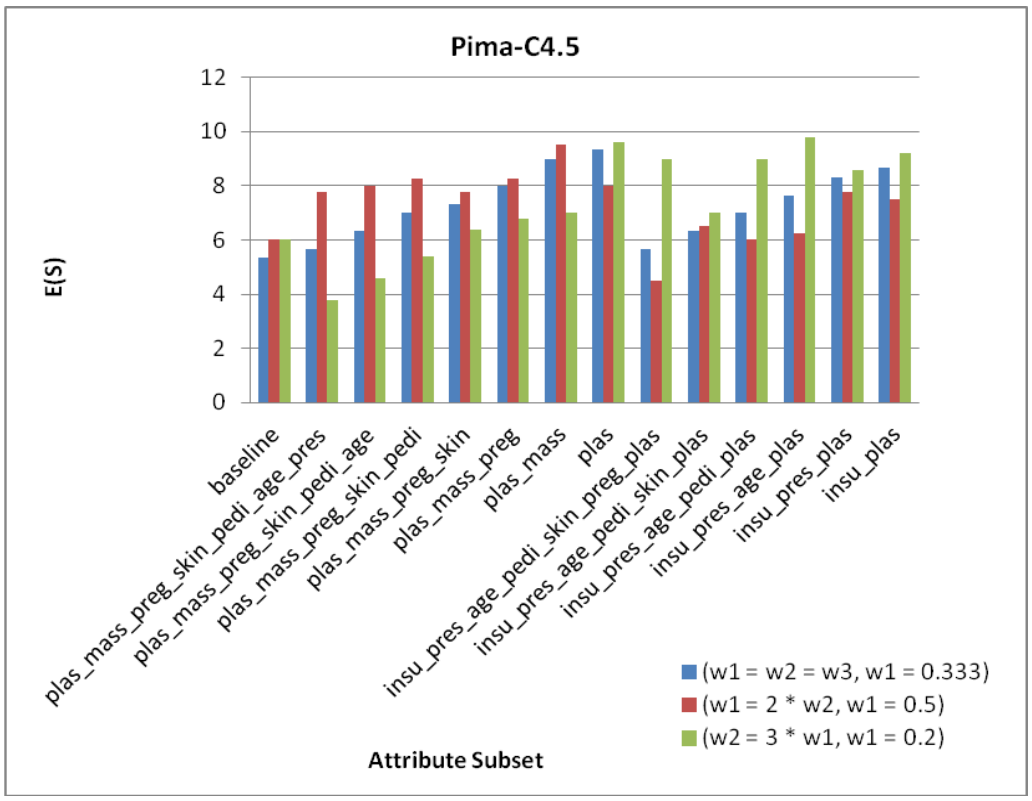

Figure 9: The E(S) corresponding to different weight ratios with respect to candidate subsets (Pima-C4.5).

From this diagram, when $w1 = w2 = w3$, subset {plas} is the selected. When $w1 = 2 * w2$, subset {plas, mass} is selected, and when $w2 = 3 * w1$, subset {insu, pres, age, plas}is selected.

6 Experiments

We used four datasets from the UCI repository. Two relatively large datasets include the *Adult* dataset (consists of 45,222 records) and the *Diabetes* dataset (consists of 101,766 records). We also considered two smaller datasets i.e., the *Pima* dataset and the *Liver Patient* dataset. The records with missing attribute values were eliminated. We obtained the results using two classification algorithms, namely, C4.5. and N.B. 10-fold cross validation was performed to evaluate the results and a statistical significant *t*-test was used to compare the results. We used Weka[1] and R[2] in order to conduct our experiment. Weka is a

[1] http://www.cs.waikato.ac.nz/ml/weka/

collection of machine learning algorithms that are used for data mining tasks. R is a language and environment for statistical computation and graphics.

We consider both cases of C!=SA and C=SA. For the case of C!=SA a PBI(DS) vs. Perf(DS) plot is obtained and the four regions discussed in Section .are shown 3.3We obtain the list of candidate attribute subsets and identify the location of each selected attribute subset within the PBI(DS) vs. Perf(DS) diagram. It is then up to the DH to choose appropriate α and identify the legitimate attribute subset from the list of available attribute subsets. It might be the case that DH provides a list of features (not the dataset) and the corresponding classification accuracy and communicates those to DR. DH only provides those attributes that satisfy the privacy requirements.

For the case of C=SA the evaluation function is implemented and the corresponding *E*(S) values for each combination of the weights is obtained. The case of Pima-C4.5 was discussed in details earlier in Section It was mentioned that, the proportion of w .5eights, especially between *w*1 and *w*2, is the choice of the DH. Having proportional relations between weights such as w1 = 4 * w2 , w2= 7 * w1 is one way of representing the relative weights. However, it is possible to consider any value for each of the weights as long as their sum is equal to 1.0. For instance the DH might decide to consider 0.34, 0.42, and 0.24 for *w*1, *w*2, and *w*3 respectively.

Let us consider the results associated with Pima-N.B. case. After obtaining *PerfR* and *PrivR* and generating the candidate-subsets, we record the *Perf*(*S*), *PBI*(*S*), and *Num*(*S*) of each of these subsets and rank them based on the methodology discussed in Section The final results are .5 shown inFigure 10.

From Figure 10, when *w*1= *w*2 = *w*3, subset {plas} is selected. When *w*1=2**w*2, and *w*2 = 0.333, subset {plas, mass, pedi, preg} is selected. When *w*2=3**w*1, and *w*1 = 0.2, subset {plas} is selected. We could have selected subset {skin, pedi, plas} or {pedi, plas} since they have the same E(S) value of 9.8. However, in such cases where there is a tie (and more than one subset have the maximum *E*(*S*) associated with them) , we may select the subset with least number of attributes i.e., {plas}.

We consider the Adult dataset next. In this case, different proportions of the weights are assumed. The corresponding results are shown in Figure 11.

2 https://www.r-project.org/

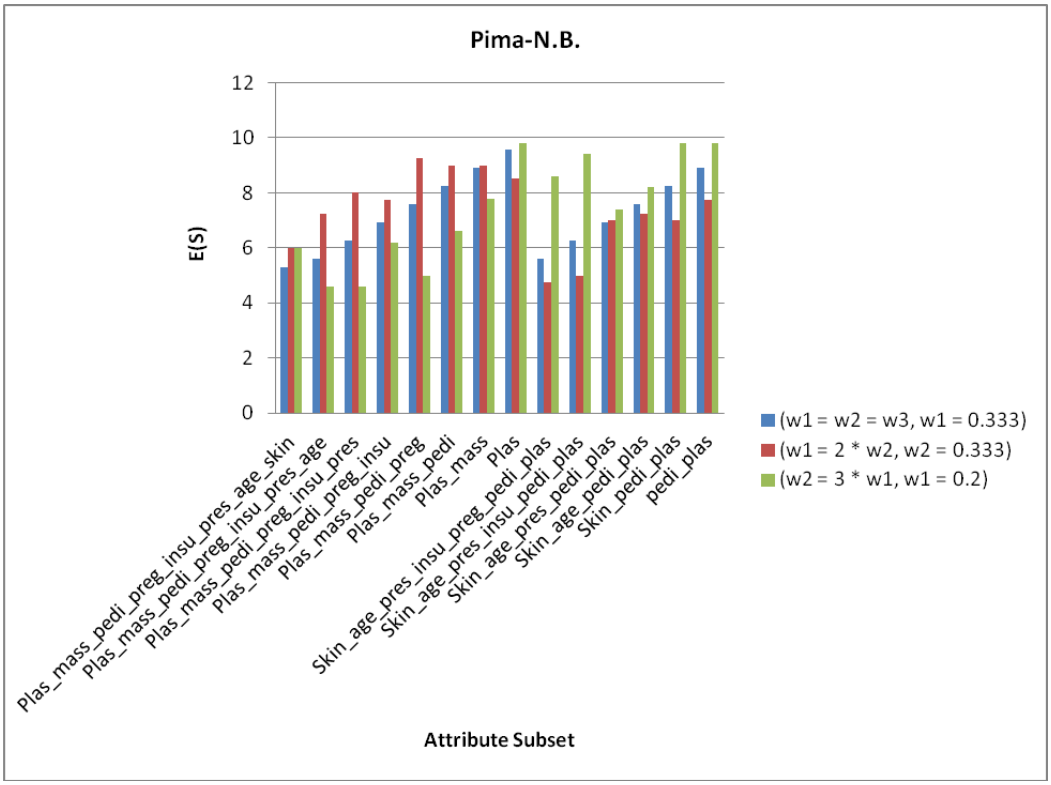

Figure 10: The E(S) corresponding to different weight ratios with respect to candidate subsets (Pima-N.B.).

Depending on the chosen weights for performance, privacy, and number of attributes, different values of E(S) per candidate subset are obtained. We record the E(S) corresponding to **argmax**($E(S)$) | $w1$, $w2$, $w3$) associated with each case (i.e., selected weight ratio) and list the results along best subset in Table 7.

Following the same methodology, similar observations are made in the case of Adult-N.B. and the Diabetes and Liver Patients datasets (both case of N.B. and C4.5 classifiers). In each case for different selected weights the corresponding E(S) is obtained.

Once again, it is possible to select an attribute subset with the best E(S) according to a given selected weighting i.e., to obtain **argmax**($E(S)$) | $w1$, $w2$, $w3$). For instance, consider the case where $w1$, $w2$, and $w3$ are 0.15, 0.3, and 0.55 respectively. In such case, the best E(S) corresponds to the subset {CG, CL, EN, RE} which implies that given those weights it represents the best selected subset. Similarly, with the same weights the second best attribute subset would correspond to {CG, CL, EN, AG, RE}.

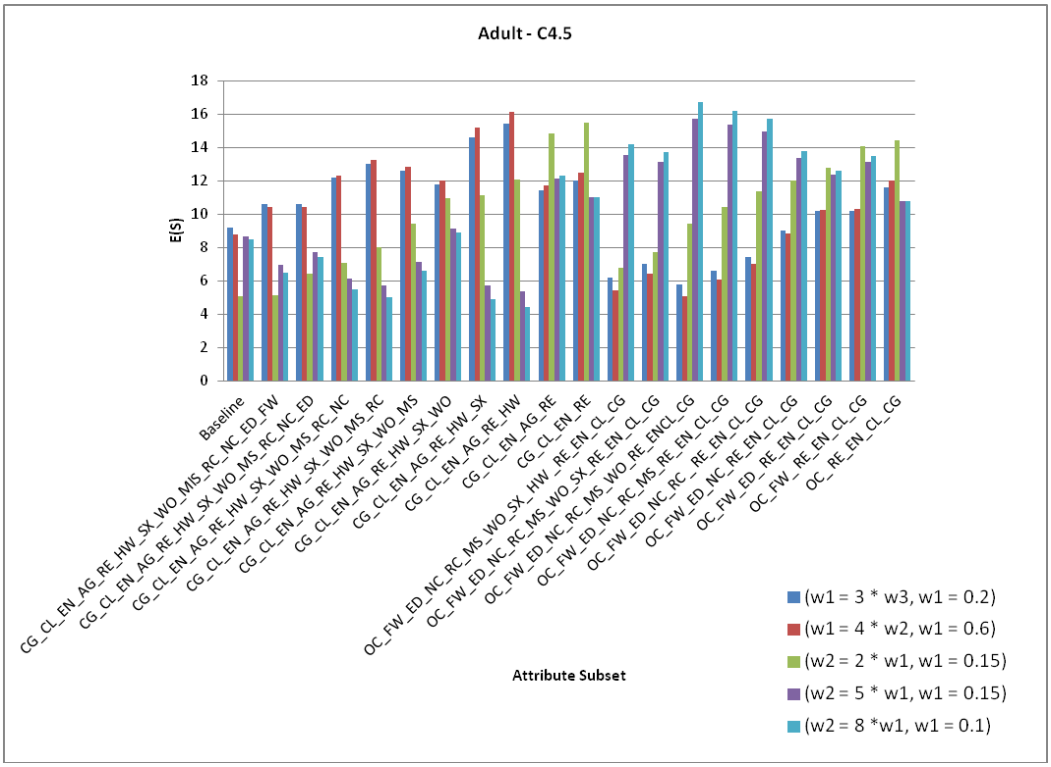

Figure 11: The E(S) corresponding to different weight ratios with respect to candidate subsets (Adult – C4.5).

We conducted further experiments with Adult dataset (case of N.B. classifier) and other datasets i.e., Diabetes and Liver Patients datasets (both case of N.B. and C4.5 classifiers). The results are shown in Appendix *A*.

In general, the results show that by introducing the *E*(S) evaluation function it is possible to identify a single attribute subset based on the DH preferences. In each case, with different combinations of privacy, utility, and dimensionality weighting it becomes possible to make decisions about the accepted/desired trade-off between these three factors, especially the privacy and utility factors. It should be noted that the proposed evaluation function is generalizable. That is, by setting any of the weights (or a combination of two of them) to zero, it is still possible to use the measure to evaluate the attribute subsets. All possible combinations are listed in Table 8. x, y, z are the weights chosen by the DH.

Table 7: Best f(E(S)) and its corresponding selected attribute subset given the weight selected by the DH.

Weights			Best Subset	E(S)
*w*1	*w*2	*w*3		
0.60	**0.20**	**0.20**	**CG, CL, EN, AG, RE, HW**	**15.4**
0.60	**0.15**	**0.25**	**CG, CL, EN, AG, RE, HW**	**16.5**
0.15	**0.30**	**0.55**	**CG, CL, EN, RE**	**15.5**
0.15	**0.75**	**0.1**	**CG, CL, EN, RE, WO, MS, RC, NC, ED, FW, OC**	**16.7**
0.1	**0.8**	**0.1**	**CG, CL, EN, RE, WO, MS, RC, NC, ED, FW, OC**	**15.75**

Table 8: Possible combinations when one weight or two weights are set to zero.

*w*1	*w*2	*w*3	*E*(*S*)
0	**0**	**1**	***w3.num* (or *num*)**
0	**1**	**0**	***w2.prv* (or *prv*)**
1	**0**	**0**	***w1.prf* (or *prf*)**
0	**x**	**y**	***w2.prv + w3.num***
x	**0**	**y**	***w1.prf + w3.num***
x	**y**	**0**	***w1.prf + w2.prv***
x	**y**	**z**	***w1.prf + w2.prv + w3.num***

From Table 8 we notice that, $E(S) = prf$ is a special case where the evaluation is based on the performance factor only. Such $E(S)$ (which depends only on the performance (e.g., accuracy)), reminds us of the evaluation used in regular non privacy-aware wrappers. In wrappers, the impact of addition/removal of a given attribute is evaluated by calculating the accuracy of the resulting attribute subset compared with accuracy of the preceding subset.

Here, the main question is, how to choose α such that it provides enough privacy. In selecting α, the very characteristics of the dataset should be taken into consideration. In general, any value of $\alpha < 0$ (i.e., being in regions III and IV) indicates more privacy when compared with baseline's PBI i.e., 0. Controlling the amount of privacy is ultimately the choice of the DH.

Given the desired weight of performance, privacy, and number of attributes, the best subset that results in the highest $E(S)$ among candidate subsets is selected and returned by the privacy-aware feature selection system that employs the proposed evaluation function. We selected *PBI* as the privacy measure associated with each selected subset of attributes. However, the measure of privacy, in such a solution could be any innate measure since our combinatory evaluation measure considers the rank of the privacy factor (*prv*) among possible candidates independent of the specific privacy measure that is used. The same argument is made about the performance measure. We consider accuracy in our experiments but it could be any other performance measure such as AUC, precision, etc. Once again, what $E(S)$ is concerned about is the rank of the performance factor (*prf*) with respect to the other candidate subsets.

This evaluation function could be applied even in the case where the original dataset is anonymized and when feature selection is applied on that anonymized data.

Recall that $E(S)$ is not tied to a particular utility/privacy measure. It is mainly used for comparison purposes in order to compare candidate subsets and to select the one with maximum value. We just mentioned that by setting the weights of one or two factors to zero, the measure still could be used to evaluate the subsets based on the remaining factor(s). The flexibility of the proposed measure also implies that we could easily expand/extend $E(S)$ to include more factors if necessary.

In addition to obtaining the maximum $E(S)$ according to chosen weights, we can also get a ranked $E(S)$ for the same weight ratios. This gives the DH more flexibility in trading-off performance, privacy and dimensionality of the resulting dataset. For example, the DH will have the option of obtaining second best subset, third best subset, and so on.

7 Conclusions and Future Work

In this work, we introduced a multi-dimensional privacy-aware evaluation function in automatic feature selection. This work is built upon and is an extension of previous work (Jafer et al., 2015a) in order to evaluate the potential attribute subsets and obtain a single best subset according to the desired preferences of efficacy (e.g., accuracy), privacy, and dimensionality of the released dataset. With the proposed multi-dimensional evaluation function, it is possible to give proportional weights or absolute weights to each of the identified factors, namely, efficacy, privacy, and dimensionality of the resulting dataset. This enables the DH to exercise its preferences and obtain a single best subset according to a given objective.

In our current work, we apply $E(S)$ to the candidate attribute subsets in order to evaluate them and to obtain the subset with maximum $E(S)$ in a given setting. The $E(S)$ evaluation measure in its extended (complete) form i.e., $E(S)=w1.prf+w2.prv+w3.num$ could be plugged in to existing feature selection techniques. We will use $E(S)$ (on-the-fly) as the evaluation measure during forward selection or backward elimination in wrappers. In such cases, rather than obtaining the complete list of candidate privacy-aware subsets, at each step, we can compare the associated{*prf*, *prv*, and *num*} rank of the updated subset (due to adding/removing of an attribute) with the {*prf*, *prv*, and *num*} rank of the preceding subset and maintain the subset with higher $E(S)$ given the weight. This logic will be applied to both cases of forward selection and backward elimination.

A potential future work is to measure the efficiency of our proposed solution with regard to the execution time. We can further examine the practicality and the scalability of the technique in real world scenarios/datasests.

Another future direction is to develop a multi-dimensional utility measure that will deliver its results in a multi-dimensional space (or at least 2-dimensional space, privacy and efficacy (accuracy) being the two dimensions). With such a measure, it becomes possible to select the right point in this two dimensional space based on the Pareto efficiency. In other words, we develop a methodology to identify the best point (being the best subset of attributes) in which no further improvement of privacy (or efficacy) is possible without harming efficacy (or privacy) respectively. In a similar setting, such a right point could be obtained based on Multi Criterion Decision-Making (MCDM). MCDM refers to a class of methods based on the idea that, given a set of decision criteria and alternatives, what would be the best alternative.

References

Ayala-Rivera, V., McDonagh, P., Cerqueus, T., & Murphy, L. (2014). A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners. *Transactions on Data Privacy, 7*(3), 337-370.

Bisiani, R. (1992). Beam Search *Encyclopedia of Artificial Intelligence* (2 ed., Vol. 2, pp. 1467-1468): John Wiley and Sons.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering, 40*(1), 16-28. doi: http://dx.doi.org/10.1016/j.compeleceng.2013.11.024

Clifton, C., & Tassa, T. (2013). On syntactic anonymity and differential privacy. *Transactions on Data Privacy, 6*(2), 161-183.

Cormode, G., Procopiuc, C. M., Shen, E., Srivastava, D., & Yu, T. (2013). *Empirical privacy and empirical utility of anonymized data.* Paper presented at the Data Engineering Workshops (ICDEW), 2013 IEEE 29th International Conference on.

Dankar, F. K., & El Emam, K. (2013). Practicing Differential Privacy in Health Care: A Review. *Transactions on Data Privacy*, *6*(1), 35-67.

Dwork, C. (2006). Differential Privacy. In M. Bugliesi, B. Preneel, V. Sassone & I. Wegener (Eds.), *Automata, Languages and Programming* (Vol. 4052, pp. 1-12): Springer Berlin Heidelberg.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. In S. Halevi & T. Rabin (Eds.), *Theory of Cryptography* (Vol. 3876, pp. 265-284): Springer Berlin Heidelberg.

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: A survey of recent developments. *ACM Comput. Surv.*, *42*(4), 1-53. doi: 10.1145/1749603.1749605

Fung, B. C., Wang, K., & Yu, P. S. (2005). *Top-Down Specialization for Information and Privacy Preservation*. Paper presented at the Proceedings of the 21st International Conference on Data Engineering.

Fung, B. C. M., Wang, K., Fu, A. W.-C., & Philip, S. Y. (2010). *Introduction to privacy-preserving data publishing: concepts and techniques*: CRC Press.

http://archive.ics.uci.edu/ml/. UCI repository. 2013

Iyengar, V. S. (2002). *Transforming data to satisfy privacy constraints*. Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, Edmonton, Alberta, Canada.

Jafer, Y. (2014). Task Oriented Privacy (TOP) Technologies. In M. Sokolova & P. van Beek (Eds.), *Advances in Artificial Intelligence* (Vol. 8436, pp. 375-380): Springer International Publishing.

Jafer, Y., Matwin, S., & Sokolova, M. (2014a). *Privacy-aware Filter-based Feature Selection.* Paper presented at the First IEEE International Workshop on Big Data Security and Privacy (BDSP 2014), Washington DC, USA.

Jafer, Y., Matwin, S., & Sokolova, M. (2014b). Task Oriented Privacy Preserving Data Publishing Using Feature Selection. In M. Sokolova & P. van Beek (Eds.), *Advances in Artificial Intelligence* (Vol. 8436, pp. 143-154): Springer International Publishing.

Jafer, Y., Matwin, S., & Sokolova, M. (2014c). Using Feature Selection to Improve the Utility of Differentially Private Data Publishing. *Procedia Computer Science*, *37*(0), 511-516. doi: http://dx.doi.org/10.1016/j.procs.2014.08.076

Jafer, Y., Matwin, S., & Sokolova, M. (2015a). *A Framework for A Privacy-aware Feature Selection Measure.* Paper presented at the Privacy, Security, and Trust (PST), Izmir, Turkey.

Jafer, Y., Matwin, S., & Sokolova, M. (2015b). Privacy-aware Wrappers. In D. Barbosa & E. Milios (Eds.), *Advances in Artificial Intelligence* (Vol. 9091, pp. 130-138): Springer International Publishing.

Janecek, A., W. N. Gansterer, M. Demel and G. Ecker (2008). On the Relationship Between Feature Selection and Classification Accuracy. JMLR: Workshop and Conference Proceedings 4: 90-195.

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006). *Workload-aware anonymization*. Paper presented at the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, Philadelphia, PA, USA.

Li, N., Li, T., & Venkatasubramanian, S. (2007). *t-Closeness: Privacy Beyond k-Anonymity and l-Diversity*. Paper presented at the proceedings of the twenty third international conference on data engineering Istanbul, Turkey.

Machanavajjhala, A., Gehrke, J., Kifer, D., & Venkitasubramaniam, M. (2006). *L-diversity: privacy beyond k-anonymity*. Paper presented at the Proceedings of the 22nd International Conference on Data Engineering, Atlanta, Georgia, US.

Nguyen, H. H., Kim, J., & Kim, Y. (2013). Differential Privacy in Practice. *Journal of Computing Science and Engineering, 7*(3), 177-186.

Powell, W. B. (2007). Approximate Dynamic Programming: Solving the curses of dimensionality, John Wiley & Sons.

Rich, E., & Knight, K. (1991). *Artificial intelligence* (2nd ed.). New York: McGraw-Hill.

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence : a modern approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Sweeney, L. (2002). Achieving k-anonymity privacy protection using generalization and suppression. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst., 10*(5), 571-588. doi: 10.1142/s021848850200165x

Tsamardinos, I., & Aliferis, C. F. (2003). *Towards principled feature selection: Relevancy, filters and wrappers.* Paper presented at the Proceedings of the ninth international workshop on Artificial Intelligence and Statistics.

Wang, K., Yu, P. S., & Chakraborty, S. (2004). *Bottom-up generalization: a data mining solution to privacy protection*. Paper presented at the Proceedings of the fourth IEEE International Conference on Data Mining, Bringhton, UK.

Wilt, C. M., Thayer, J. T., & Ruml, W. (2010). *A comparison of greedy search algorithms.* Paper presented at the Third Annual Symposium on Combinatorial Search.

Yu, L., & Liu, H. (2004). Efficient Feature Selection via Analysis of Relevance and Redundancy. *J. Mach. Learn. Res., 5*, 1205-1224.

Appendix A: The E(S) plots Corresponding to Different Datasets

The following E(S) plots correspond to different dataset/algorithm combinations. In each case, the E(S) corresponding to different weight ratios with respect to candidate subsets is shown. Depending on the selected ratio, a best subset is identified.

Figure 12 shows the E(S) corresponding to Adult-N.B. Depending on the w1, w2, and w3 weights, the E(S) of a given subset is calculated. Furthermore, we can identify the best E(S) given the weight combinations. For example, when w1, w2, and w3 weights are 0.6, 0.15, and 0.25 respectively, the best subset corresponds to {CG, RE, MS, OC, ED, EN, AG, HW, SX, NC, WO, RC}.

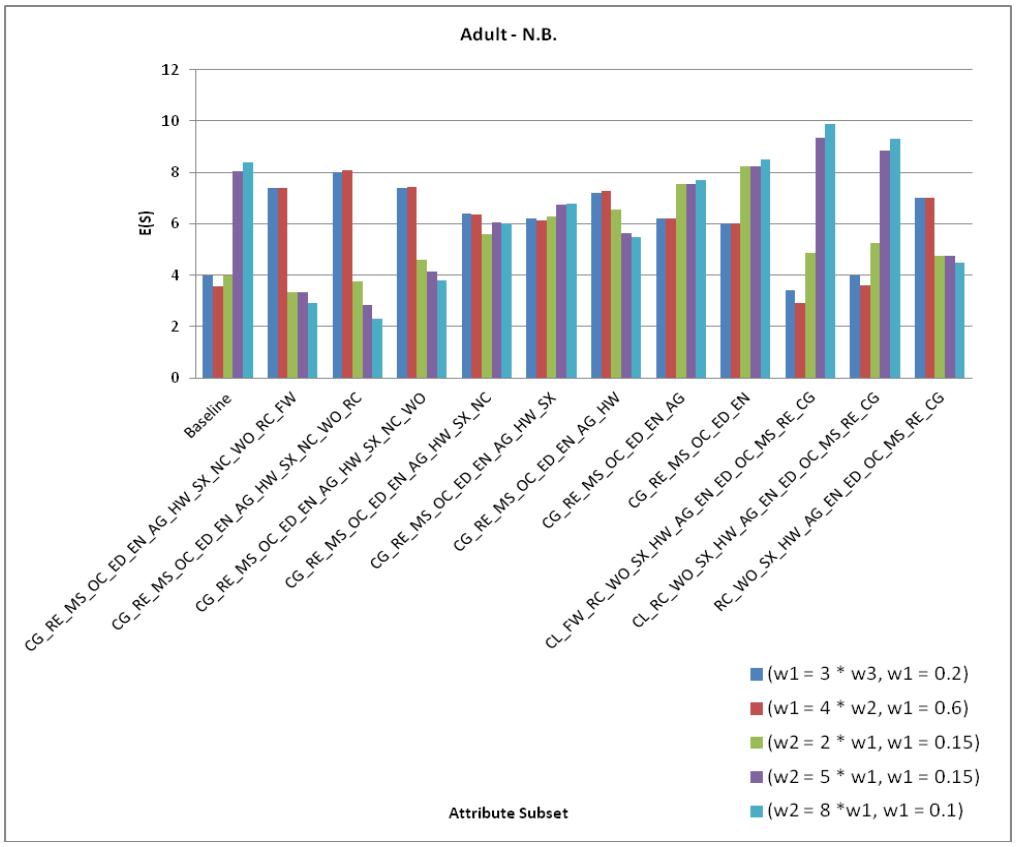

Figure 12: The E(S) corresponding to different weight ratios with respect to candidate subsets (Adult – N.B.).

Figure 13 shows the E(S) corresponding to the Diabetes dataset and the C4.5 classifier. When w1, w2, and w3 are 0.25, 0.5, and 0.25, subset {NI, DDID, DI, D3, Age, ME, ASID, Insu} gives the highest E(S) score. For the same weights, however, the lowest E(S) score is associated with {NI, DDID, Age, Me, ASID, Insu}.

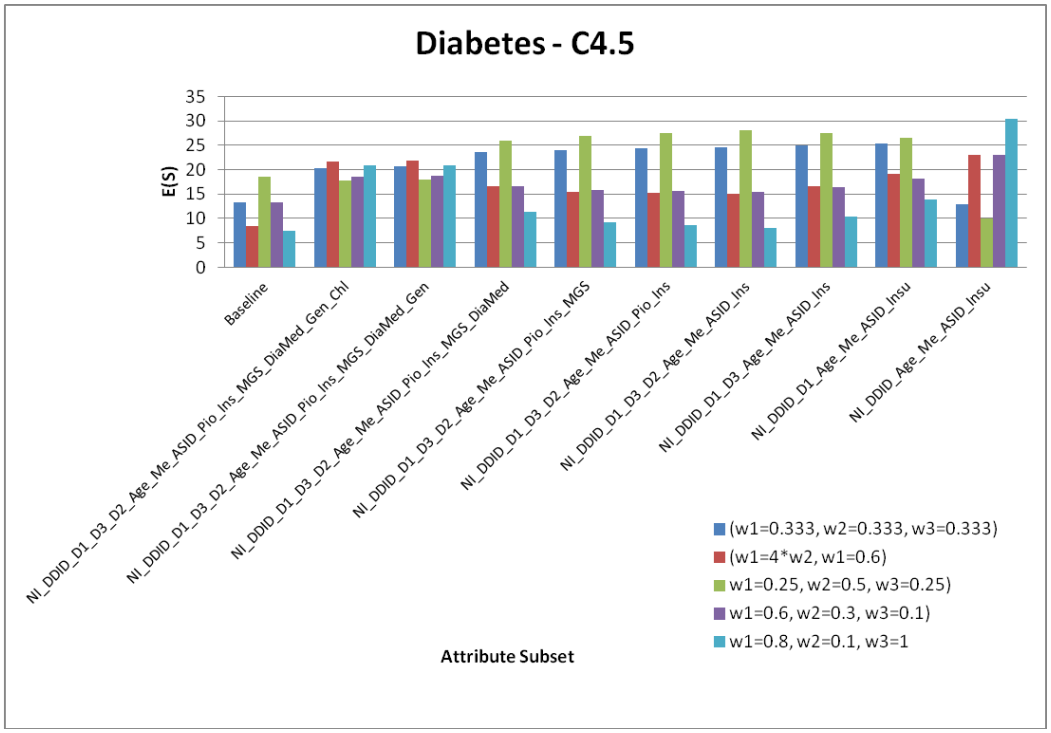

Figure 13: The E(S) corresponding to different weight ratios with respect to selected candidate subsets (Diabetes - C4.5).

Figure 14 corresponds to Diabetes – N.B. When the w1, w2, and w3 weights are 0.6, 0.15, and 0.25 respectively, subset {NI, DI, NE, D2, DDID, ND, Glim, Age, Gen, DiaMed, D3} results in the best E(S) score.

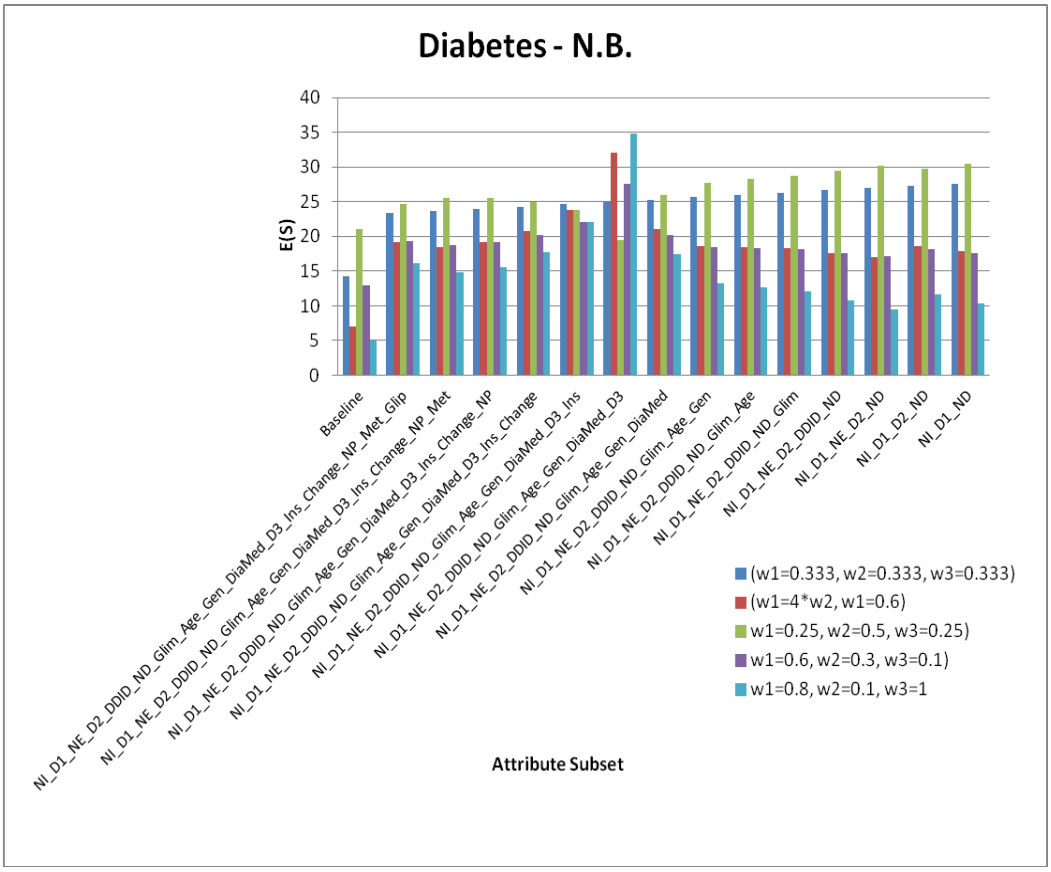

Figure 14:The E(S) corresponding to different weight ratios with respect to selected candidate subsets (Diabetes - N.B.).

In the case of Liver Patients – C4.5, The best E(S) corresponds to the subset {AG, TP, SAIA} when the w1, w2, and w3 weights are 0.8, 0.1, and 0.1 respectively. The same subset has the best E(S) when the weights are 0.6, 0.15, and 0.25 respectively.

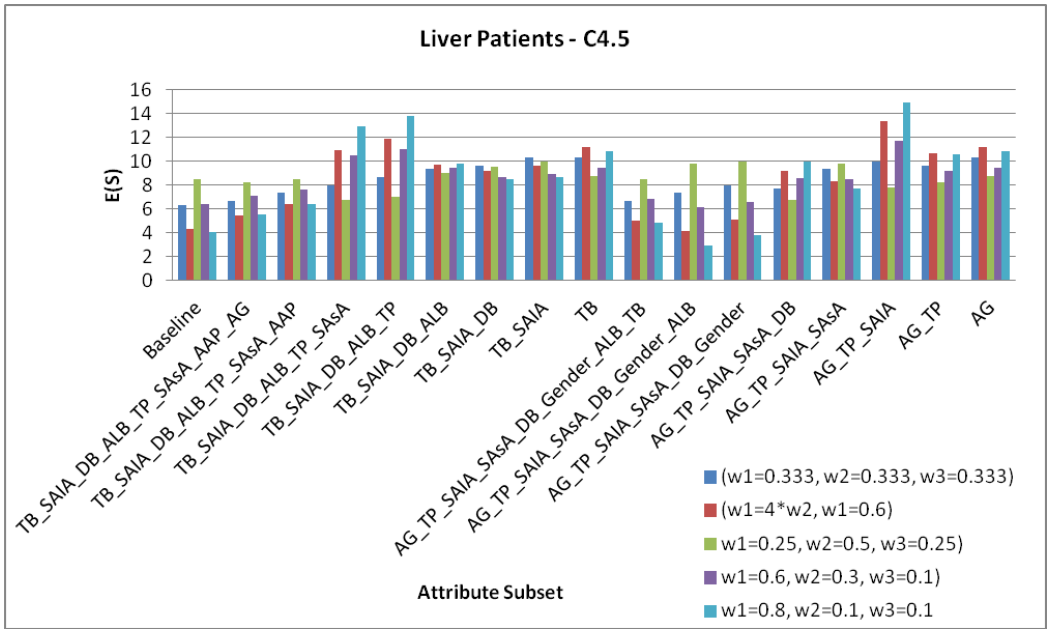

Figure 15: The E(S) corresponding to different weight ratios with respect to candidate subsets (Liver Patients - C4.5).

Figure 16 refers to the case of LiverPatients – N.B. When the w1, w2, and w3 weights are 0.8, 0.1, and 0.1, the best E(S) is associated with the subset {AAP, SAIA, Gender, AG, ALB, TP, DB}.

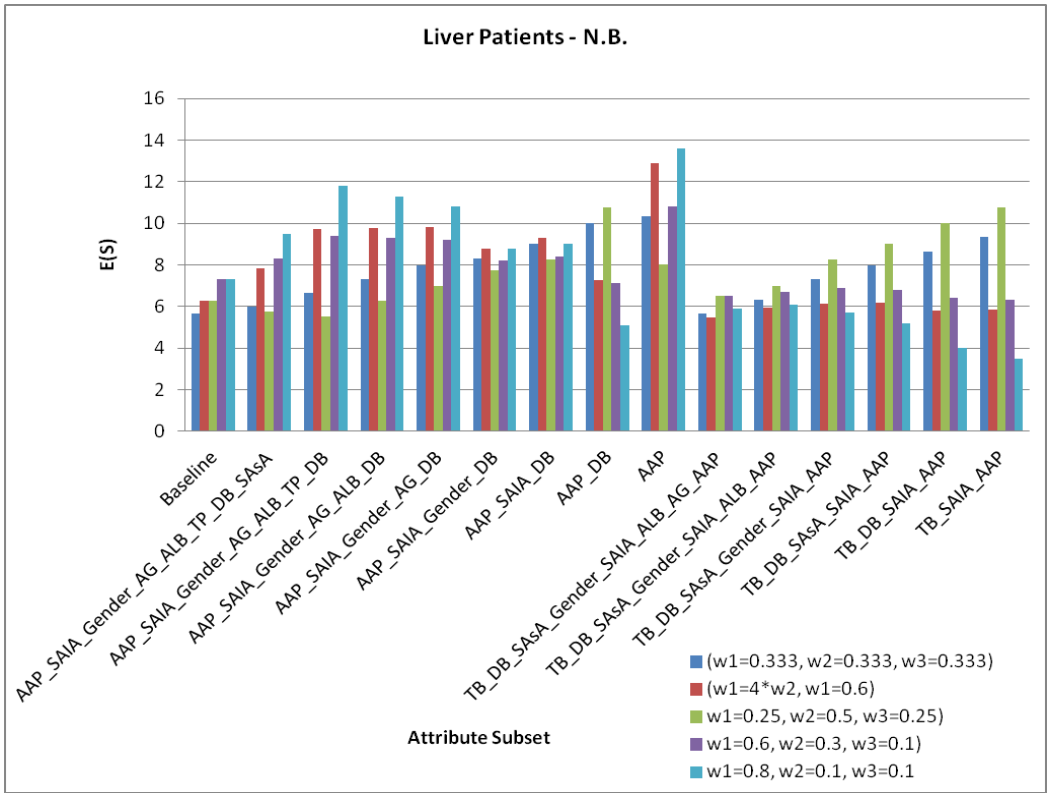

Figure 16: The E(S) corresponding to different weight ratios with respect to candidate subsets (Liver Patients - N.B.).