View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by CiteSeerX

An Evolutionary Optimization Approach for Categorical Data Protection

Jordi Marés
Artificial Intelligence Research Institute (IIIA)
Spanish Council of Scientific Research (CSIC)
Campus de la UAB, s/n
Bellaterra, Spain
jmares@iiia.csic.es

Vicenç Torra
Artificial Intelligence Research Institute (IIIA)
Spanish Council of Scientific Research (CSIC)
Campus de la UAB, s/n
Bellaterra, Spain
vtorra@iiia.csic.es

ABSTRACT

The continuous growing amount of public sensible data has increased the risk of breaking the privacy of people or institutions in those datasets. Many protection methods have been developed to solve this problem by either distorting or generalizing data but taking into account the difficult trade-off between data utility (information loss) and protection against disclosure (disclosure risk).

In this paper we present an optimization approach for data protection based on an evolutionary algorithm which is guided by a combination of information loss and disclosure risk measures. In this way, state-of-the-art protection methods are combined to obtain new data protections with a better trade-off between these two measures. The paper presents several experimental results that assess the performance of our approach.

Keywords

Genetic algorithms, data privacy, categorical data, data mining, information loss, disclosure risk

1. INTRODUCTION

Statistical agencies and other institutions are supposed to protect the confidentiality of the people or entities when a dataset is published. However this task of protecting a dataset is not as trivial as just removing any explicit individual's identifier such as Social Security Numbers, or entity names (see [1] for details).

Different approaches have been developed to obtain better protections [9], and they are divided usually into two groups. The methods in one are called *perturbative methods* and are the ones that perform a data distortion changing values by others that may not be related with their meaning. The main representative protection methods in this group are Microaggregation [7], Rank Swapping [14], and Post Randomization Method (PRAM) [4]. The methods in the other group are called *non-perturbative methods* and are the ones that instead of changing the meaning of some values, they try to generalize or suppress the values without loosing the essential meaning. The main non-perturbative protection methods are: Top Coding, Bottom Coding, and Global Recoding [6].

However, these protection methods cannot ensure the privacy maintaining the utility. For that reason the quality of the protection must be assessed by dealing with two competing goals: the micro-data file has to be safe enough to guarantee the protection of individual respondents but at the same time the loss of information should not be too large. The discussion for this can be found in [3].

Datasets might contain different types of data. Two of the most frequent ones are continuous data and categorical data. The protection of continuous data can take advantage of the number of possible values a variable can take, and the arithmetic operations on them. In the categorical case, the options are more limited. The number of categories is usualy small, almost no operations exist on these variables, and terms might have a rellevant semantics that needs to be kept in the protection process. Then, the actions that can be performed with categorical data are almost restricted to an exchange of categories by others that already exist, and a generalization of some categories into more general ones. The limitation on the possible actions makes protection a difficult task. In this work, we precisely focus on this categorical data protection case.

In this paper, we show how information loss and disclosure risk can be integrated within an evolutionary algorithm to seek new and enhanced protections for categorical data. We propose an approach that permits us to combine state-of-the-art protection methods with a post-masking evolutionary algorithm optimization.

The reminder of the paper is organized as follows. Section 2 describes our evolutionary approach as well as they main operators. Several experimental results are shown in Section 3. Finally, Section 4 contains concluding remarks and possible future work.

2. EVOLUTIONARY ALGORITHM TO ENHANCE DATA PRIVACY

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
PAIS 2012, March 30, 2012, Berlin, Germany.
Copyright 2012 ACM 978-1-4503-1143-4/12/03 ...$10.00

Evolutionary algorithms are stochastic processes oriented to find exact or approximate solutions to optimization or search problems. Those algorithms have a population of individuals, denoted as $P(t)$ for generation t, where each individual $X_i \in P(t)$ is associated to a potential solution of a given problem. In order to guide the individuals from generation to generation a fitness evaluation function is used. And, finally, some of the selected individuals are altered by operators with an evolutive connotation, such as mutation and crossover, which generate new offsprings with chances to be inserted into the population as new individuals.

Then, the main idea behind this work is to use the state-of-the-art protection methods for categorical data together with this kind of algorithms in order to obtain good protections for a specific file just combining pairs of protected files or altering their values in an evolutive way. The initial population of our algorithm is composed by several protected files, so each individual is a specific protection of the same file, and the main goal is to find the protected file that best satisfies our fitness function.

Algorithm 1 shows our approach. Recall that this can be seen as an instantiation of a generic evolutionary algorithm, with some particularities described below.

Algorithm 1 Evolutionary Algorithm to Enhance Privacy.

Input: $P(0) = \{X'_i\}$ initial population of protections for X.
Output: $P(t) = \{X'_j\}$ generation t.
$t \Leftarrow 0$
evaluate($P(0)$)
while stopping($P(t)$) $\neq$ `true`; **do**
  alter $\Leftarrow$ randomly choose between mutation and cross
  if alter by mutation **then**
    $i \Leftarrow$ select(N)
    $X'_j \Leftarrow$ mutate(X'_i)
    evaluate(X'_i, X'_j)
  else
    $\{i_1, i_2\} \Leftarrow$ select(N_b, N)
    $X'_{j_1}, X'_{j_2} \Leftarrow$ cross(X'_{i_1}, X'_{i_2})
    evaluate($X'_{i_1}, X'_{i_2}, X'_{j_1}, X'_{j_2}$)
  end if
  $t \Leftarrow t + 1$
end while
return $P(t)$

All the key points of Algorithm 1 are described in the following subsections. Subsection 2.1 describes the individual representation, Subsection 2.2 describes the genetic operators, Subsection 2.3 describe the evaluation function, and 2.4 describes the method of selecting and replacing individuals from the population.

2.1 Genotype Encoding

Usually, evolutionary algorithms are applied to continuous numerical data where the values can be converted into a binary numbers and then it is possible to change their meaning (values) by altering bits. This is done in this way to avoid abrupt changes in the values.

However, in this paper we are dealing with categorical data and this is a kind of data that only have meaning in the form of a string and, in addition, its meaning can only be modified by changing the entire string, so partial modifications of the string can generate categories out of our domain.

For that reason we decided to deal with the original categories directly without any type of encoding, this is, the chromosomes of method's population are just the protected data-files read and loaded into memory, where the genes are the string values. In this way, the space complexity of our approach will be determined by the number of registers n in the data, the number of attributes a, and the number of files loaded into memory f, obtaining a space complexity of $O(n * a * f)$.

2.2 Genetic Operators

Our proposed algorithm is based on two basic genetic operators: mutation, and crossover [12]. Both crossover and mutation rate are chosen heuristically, and we have decided to use a rate of 0.5 for both. A random value (*alter*) between 0 and 1 decide which operation is going to be executed, using the value 0.5 as a delimiter.

2.2.1 Mutation

In the case of continuous data people use mutation in order to randomly alter bits of the genes to obtain a new offspring. In our case with categorical data, we also want to obtain a new offspring but we can not alter some parts of the values randomly. In addition, we have also to deal with the constraint of that each variable have a limited number of categories admitted as a valid values, and we need to take it into account when altering them.

So, we decided to define the mutation operation as follows. Given a chromosome X (i.e. a protected datafile), it is mutated by randomly selecting a gene x_i (i.e. a string value) and changing it by a randomly selected value among all valid values for the specific variable v_i.

This operator is represented in Algorithm 1 as the *mutate* function.

2.2.2 Crossover

Crossover is a genetic operator that consists on recombine values from two chromosomes obtaining also two new offsprings. In our case, crossover of two masked datasets X and Y is performed by a 2-point crossing at the category level as follows. Take a value position s at random as the first point, and consider that the two values at this position are $x_s \in X$ and $y_s \in Y$. Take another value position r at random in the range $[s, length(X) - 1]$, and consider the two values at this position are $x_r \in X$ and $y_r \in Y$.

When $s = r$ there is only one value selected, so only this value will be swapped obtaining two new offsprings $Z_1 = \{x_1, \ldots, x_{s-1}, y_s, x_{s+1}, \ldots, x_n\}$ and $Z_2 = \{y_1, \ldots, y_{s-1}, x_s, y_{s+1}, \ldots, y_n\}$.

When $s \neq r$ all values between the two points have to be swapped obtaining two new offsprings $Z_1 = \{x_1, \ldots, x_{s-1}, y_s, y_{s+1}, \ldots, y_r, x_{r+1}, \ldots, x_n\}$ and $Z_2 = \{y_1, \ldots, y_{s-1}, x_s, x_{s+1}, \ldots, x_r, y_{r+1}, \ldots, y_n\}$.

This operator is represented in Algorithm 1 as the *cross* function.

2.3 Fitness Function

In order to evaluate the degree of protection of each individual we use the two most widely used measures in data privacy, the information loss and the disclosure risk, and then they are aggregated into a single score value.

2.3.1 Information Loss

Information loss [8] measures the quantity of harm that is inflected to the data when it is protected. This measure is small when the analytic structure of the masked dataset is very similar to the structure of the original dataset, so, the motivation for preserving the structure of the dataset is to ensure that the masked dataset will be analytically valid and interesting.

Information loss measures used in this work are: contingency table-based information loss (CTBIL) [8], distance-based information loss (DBIL) [8], and entropy-based information loss (EBIL) [15]. Then, the final information loss result is the average of all three measures.

2.3.2 Disclosure Risk

It is not enough to assess the protection quality with only the measurement of information loss, disclosure risk [2] is also needed. Disclosure risk measures are to evaluate in what extend some information can be obtained about the individuals from the protected data set. Different approaches exist about the meaning of disclosure risk. In this case, we follow the approach based on identity disclosure. That is, the intruder is able to link a record to a particular individual. Other approaches includes attribute disclosure where disclosure risk also includes when the intruder can improve his knowledge about a particular attribute of an individual without linking any record to this particular individual. E.g., have a rough estimation of the income of Lois Lane in Metropolis. This measure is small when the masked dataset values are different respect the original ones.

In this case, the disclosure risk measures used in this work are: interval disclosure (ID) [2], distance-based record linkage (DBRL) [16], probabilistic record linkage (PRL) [16] and, rank swapping record linkage (RSRL) [17]. Then, the final disclosure risk result is the average of all four measures.

2.3.3 Aggregated Score

At this point we have a measure for the information loss and a measure for the disclosure risk, so we are dealing with a multi-objective optimization problem. In order to address this problem we need to provide to the evolutionary algorithm a single score value which reflects the quality of the protection depending on the values of both measures. In this work we used two different score aggregation functions focusing on different aspects.

The first score function is just the mean value of information loss (IL) and disclosure risk (DR) measures, and it is shown in Equation 1.

$$Score(X) = \frac{IL(X) + DR(X)}{2} \tag{1}$$

Expression 1 has been used in several papers [2][18] . According to this expression, the best protection is achieved with a minimum value for each measure (i.e., IL=0 and DR=0). Nevertheless, given a particular score we prefer to have the same value in both scores. That is, for a score of 20%, we prefer IL=20 and DR=20 than IL=0 and DR=40. Expression 1 cannot represent this preference appropriately.

To better represent our choices, we present an alternative function that does not permit such perfect trade-off between information loss and disclosure risk. The expression is the maximum of information loss and disclosure risk (see Equation 2).

$$Score(X) = max(IL(X), DR(X)) \tag{2}$$

This second function penalizes a protected dataset that has a large unbalance between disclosure risk and information loss. Note that just one bad value of IL or DR leads to a bad score.

2.4 Selection and Replacement Methods

Consider that the current population $P(t) = \{X_1, \ldots, X_n\}$ is sorted by the score function, with $Score(X_i) \leq Score(X_j)$ whenever $i \leq j$. Given a fixed parameter N corresponding to the population size, the selection method filters the best N_b individuals in terms of its score value, and selects an individual X_i in a different way when performing mutation than when performing crossover.

Equation 3 is a probabilistic strategy which is proportional to the fitness function [10][12]. With proportional selection, better individuals have a greater probability of being selected.

$$p(X_i) = \frac{Score(X_i)}{\sum_{j=1}^{N} Score(X_j)} \tag{3}$$

In the mutation case, an individual X_i is chosen from the current entire population with probability $p(X_i)$ given by Equation 3.

From this selected individual it is obtained a potentially new individual (as described in Section 2.2 above). During the evaluation, an elitism replacement strategy is followed, which means that the two individuals are compared and only the individual with the best fitness value survives. The use of the elitism replacement strategy is to guarantee that the next generation individual will be at least not worse than the actual, then, it can prevent a loss of the best solution found.

In the crossover case, two individuals X_{i1} and X_{i2} are chosen. X_{i1} is selected randomly from a leader group with the

N_b best scores. The second individual X_{i2} is chosen from the entire population using the probabilistic method shown in Equation 3. A recombination of these two individuals produce two new individuals, X_{j1} and X_{j2} (as described in Section 2.2 above). In our case, each newcomer X_{jk} maintains a proximity relation with its parent X_{ik}. During the evaluation, an elitist niching method - known as Deterministic Crowding (DC) [11][13] - is followed such that only individuals with the best fitness value in each pair (X_{ik}, X_{jk}) survive. Should be noticed that this method is effective in maintaining the diversity of the population in terms of genotypic search space, but it does not necessarily guarantee the diversity of maskings.

3. EXPERIMENTAL RESULTS

In this section we illustrate and empirically evaluate our proposed method. In the experiments we used four different datasets extracted from [5]. The first dataset is the U.S. Housing Survey of 1993 and consists of 1000 records with 11 categorical attributes containing information about housing values of different people. The second is the German Credit dataset and consists of 1000 records with 13 categorical attributes containing information about the credit risk of German people. The third dataset is the Solar Flare dataset and consists of 1066 records with 13 categorical attributes containing information about different detected solar flares. Finally, the fourth is the Adult dataset and consists of 1000 records with 8 categorical attributes containing information about the average income of different type of people.

For each dataset we constructed a population of protections using the state-of-the-art protection techniques: Microaggregation, Bottom Coding, Top Coding, Global Recoding, Rank Swapping and, Post Randomization Method (PRAM). For the first dataset we had a population of 110 protections (72 of Microaggregation, 6 of Bottom Coding, 6 of Top Coding, 6 of Global Recoding, 11 of Rank Swapping and, 9 of PRAM). For the second and third datasets we had a population with a 104 protections for each one (72 of Microaggregation, 4 of Bottom Coding, 4 of Top Coding, 4 of Global Recoding, 11 of Rank Swapping and, 9 of PRAM). The last dataset had a population of 86 protections (48 of Microaggregation, 6 of Bottom Coding, 6 of Top Coding, 6 of Global Recoding, 11 of Rank Swapping and, 9 of PRAM).

Regarding the attributes selected to protect in each dataset are as follows. For the Housing dataset we protected three attributes: BUILT with 25 categories, DEGREE with 8 categories and, GRADE1 with 21 categories. In the case of German dataset: EXISTACC with 5 categories, SAVINGS with 6 categories, and PRESEMPLOY with 6 categories. Flare dataset attributes are: CLASS with 8 categories, LARGSPOT with 7 categories, and SPOTDIST with 5 categories. Finally, Adult dataset protected attributes are: EDUCATION with 16 categories, MARITAL-STATUS with 7 categories, and OCCUPATION with 14 categories.

To test the performance of our approach we performed three different kind of experiments explained in the following subsections. In Subsection 3.1 we present an experiment that consists of using the mean of information loss and disclosure risk as a score. Our second experiment is presented in Subsection 3.2 and, in this case, we use the max value of both measures as a score. Finally, in Subsection 3.3 we present an experiment to prove the robustness of our approach when the best initial individuals are missing.

3.1 First experiment: using mean value as a score

In this first experiment we applied our evolutionary algorithm using the fitness function that uses the mean values of both information loss and disclosure risk as a score shown in Equation 1 to all four dataset populations independently.

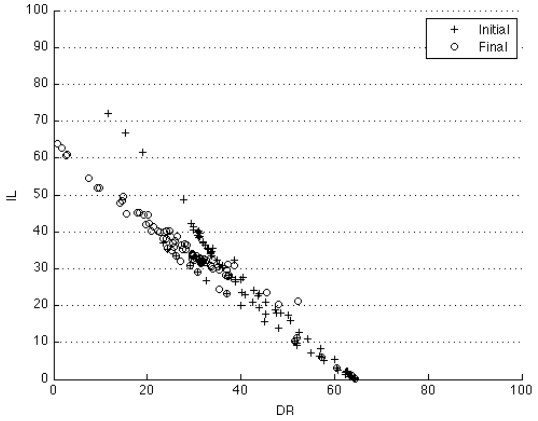

Figure 1: Dispersion plot of initial and final population information loss and disclosure risk for the Adult dataset using fitness Equation 1.

In order to evaluate the results of our experiment we splitted our analysis into two parts. The first part of our analysis is focused on the initial and final pairs of values (IL,DR) for all datasets shown in Figure 1 for the Adult dataset, Figure 3 for the Housing dataset, Figure 5 for the German dataset and, Figure 7 for the Flare dataset.

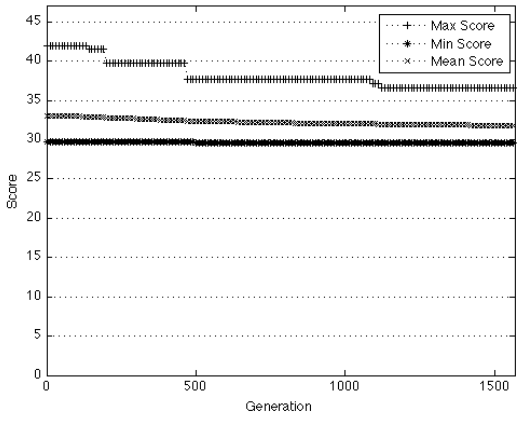

Figure 2: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the Adult dataset using fitness Equation 1.

It can be noticed that, in all the cases, final population is

more optimized than the initial population because of the reduction of the values in the tuples (IL,DR). However, there also exist individuals in the final population that have reduced their score value but obtaining an individual with very unbalanced measures. Recall that, according to our preferences, given a certain score, we prefer balanced information loss and disclosure risk. Furthermore, this effect does not appear in the same degree to all datasets. It can be seen that the Flare and German datasets have more unbalanced final individuals than the Housing and Adult datasets.

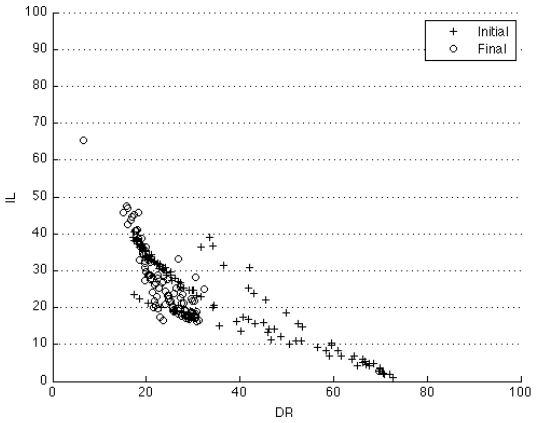

Figure 3: Dispersion plot of initial and final population information loss and disclosure risk for the Housing dataset using fitness Equation 1.

The second part of our analysis focuses on the evolution of the max, mean and min score values of the population during all the generations shown in Figure 2 for the Adult dataset, Figure 4 for the Housing dataset, Figure 6 for the German dataset and, Figure 8 for the Flare dataset.

In these figures it can be seen that max score has few decrements but most of them are quite abrupt, and this is because our selection policy gives few opportunities to the individuals with bad score to be selected, and when they are selected they almost always have a considerable improvement of their score value using parts of other better individuals. The improvements obtained for the max score are the following: in the case of the Adult dataset we had a decrement from 41.95 to 36.6 (12.75% of improvement), for the Housing dataset we obtained a decrement from 36.96 to 36.14 (2.22% of improvement), for the German dataset it was from 36.59 to 31.74 (13.25% of improvement) and, for the Flare dataset this max score decreased from 42.53 to 33.56 (21.09% of improvement).

Looking at the evolution of the mean score it can be seen that it has more or less continuous decrement and this is what we expected because in almost every iteration there is a score improvement for an individual, so the mean score of the entire population is also improved. Concretely, the improvement obtained for the mean score in all datasets during this first experiment is as follows: in the case of the Adult dataset we had a decrement from 33.05 to 31.78 (3.84% of improvement), for the Housing dataset the decrement was from 29.79 to 25.25 (15.24% of improvement), for the German dataset it was from 29.37 to 28.91 (1.57% of improvement) and, for the Flare dataset it was from 29.57 to 28.13 (4.87% of improvement).

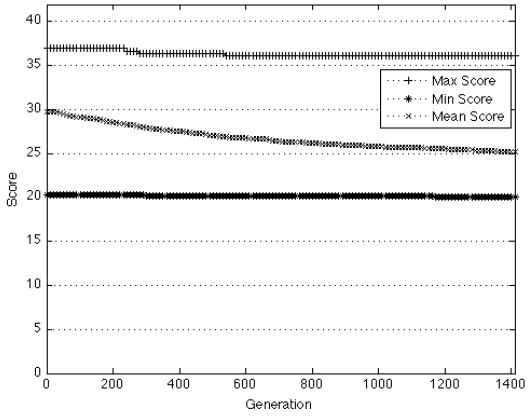

Figure 4: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the Housing dataset using fitness Equation 1.

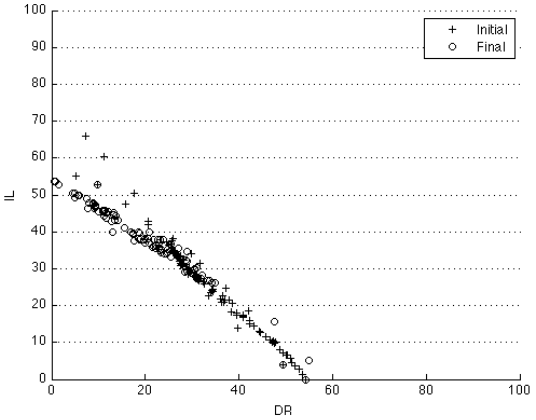

Figure 5: Dispersion plot of intial and final population information loss and disclosure risk for the German dataset using fitness Equation 1.

The last score to analyze is the min score evolution. In this case it can be noticed that the improvement is very small and the reason for this is that it is very difficult to improve a protected dataset that already has a good score (in terms of the fitness function used) using other protected files with a worse score. The improvements obtained for this min score are as follows: for the Adult dataset we obtained a decrement from 29.68 to 29.61 (0.24% of improvement), in the case of the Housing dataset there is a decrement from 20.36 to 20.12 (1.18% of improvement), for the German dataset we obtained a decrement from 26.68 to 26.54 (0.52% of improvement) and, for the Flare dataset did not obtain any decrement.

To summarize the results found after the first experiment,

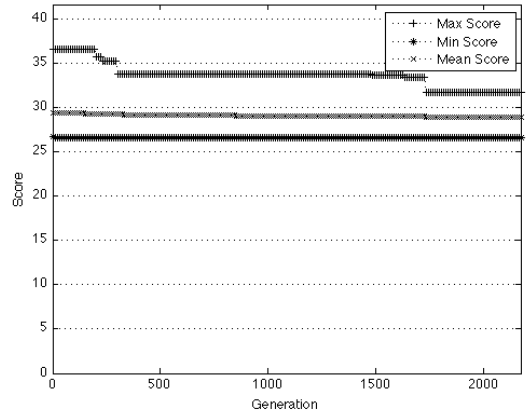

Figure 6: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the German dataset using fitness Equation 1.

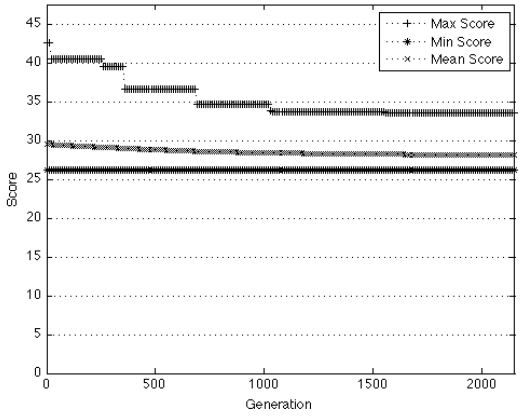

Figure 8: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the Flare dataset using fitness Equation 1.

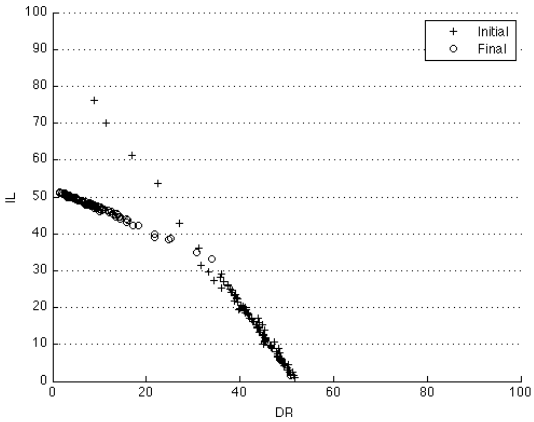

Figure 7: Dispersion plot of initial and final population information loss and disclosure risk for the Flare dataset using fitness Equation 1.

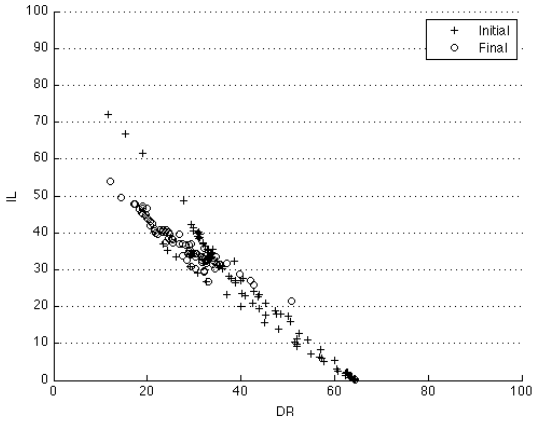

Figure 9: Dispersion plot of intial and final population information loss and disclosure risk for the Adult dataset using fitness Equation 2.

we can say that the fitness function shown in Equation 1 is not very appropriate for categorical data because it does not permit to discriminate individuals with a high unbalance and those with a low score in both measures. In addition, unfortunately, the alteration of values in categorical datasets produces quite high modifications in information loss and disclosure risk values because of the limited number of available categories to use.

3.2 Second experiment: using max value as a score

In this second experiment we wanted to try to improve the results obtained in the first experiment by applying the same evolutionary algorithm but using the fitness function shown in Equation 2 which takes as a score the maximum value between information loss and disclosure risk.

In this experiment we also splitted our analysis into two parts. The first part of our analysis is focused on the initial and final pairs of values (IL,DR) for all datasets shown in Figure 9 for the Adult dataset, Figure 11 for the Housing dataset, Figure 13 for the German dataset and, Figure 15 for the Flare dataset.

It can be seen that final population is more concentrated (in general) to pairs of (IL,DR) with more equal values than the original population (compare with Figures 1, 3, 5, and 7 of the first experiment). This was the expected behavior because the fitness function require to have low values in both measures in order to declare a new individual better than the parent.

The second part of this second experiment analysis focuses on the evolution of the max, mean and min score values of the population during all the generations. This is shown in Figure 10 for the Adult dataset, Figure 12 for the Housing dataset, Figure 14 for the German dataset and, Figure 16

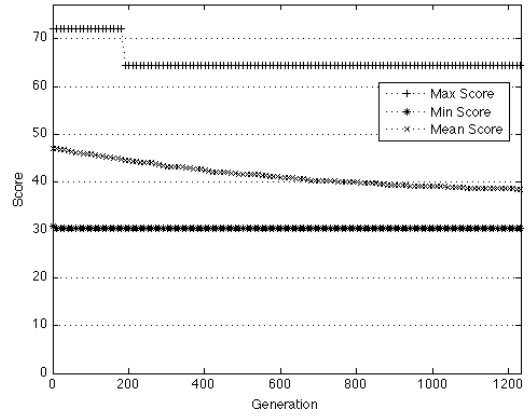

Figure 10: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the Adult dataset using fitness Equation 2.

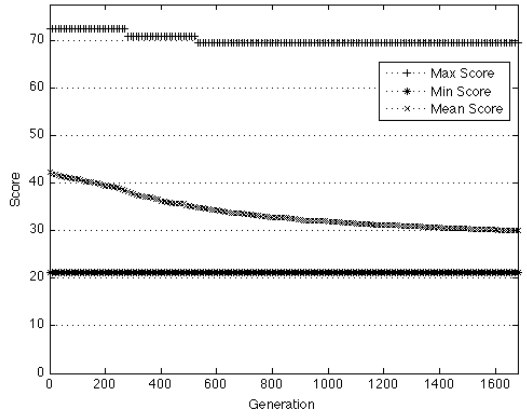

Figure 12: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the Housing dataset using fitness Equation 2.

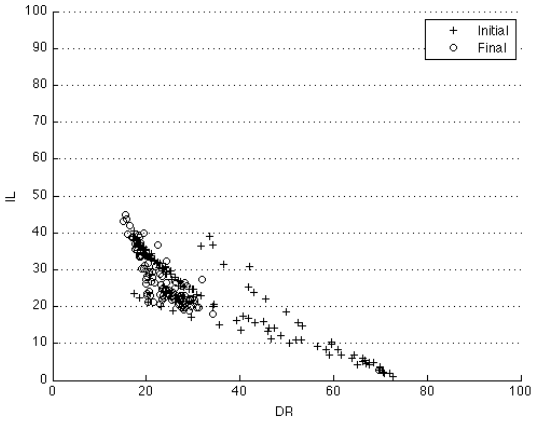

Figure 11: Dispersion plot of intial and final population information loss and disclosure risk for the Housing dataset using fitness Equation 2.

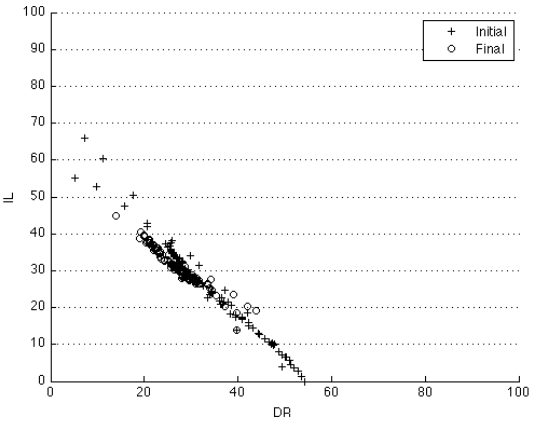

Figure 13: Dispersion plot of initial and final population information loss and disclosure risk for the German dataset using Equation 2.

for the Flare dataset.

It can be seen that max score decreases quite abruptly in some points for all datasets and remain stable between those points because our selection method gives more chances to the *best* individuals. The improvements obtained for this max score are as follows: for the Adult dataset we obtained a decrement from 72.19 to 64.38 (10.82% of improvement), in the case of the Housing dataset there is a decrement from 72.65 to 69.63 (4.16% of improvement), for the German dataset we obtained a decrement from 65.87 to 44.85 (31.91% of improvement) and, for the Flare dataset it went from 76.17 to 50.22 (34.07% of improvement).

For the mean score evolution we have that it decreases at almost every generation in all cases and its value evolves towards the value of the min score because most of the times the individuals are improved using the one with minimum score, so they go close to this min score value. In this case, the improvements obtained are as follows: for the Adult dataset we obtained a decrement from 47.05 to 38.57 (18.02% of improvement), in the case of the Housing dataset there is a decrement from 42.32 to 30.12 (28.83% of improvement), for the German dataset we obtained a decrement from 40.76 to 33.42 (18.01% of improvement) and, for the Flare dataset it went from 44.83 to 36.36 (18.89% of improvement).

Finally, the min score has little decrement in all the datasets because it is difficult to get a big improvement in this value using individuals with worse score. The improvements obtained for the min score are as follows: for the Adult dataset we obtained a decrement from 30.70 to 30.28 (1.34% of improvement), in the case of the Housing dataset there is no decrement for this score, for the German dataset we obtained a decrement from 29.18 to 28.05 (3.87% of improvement) and, for the Flare dataset it went from 31.77 to 31.63

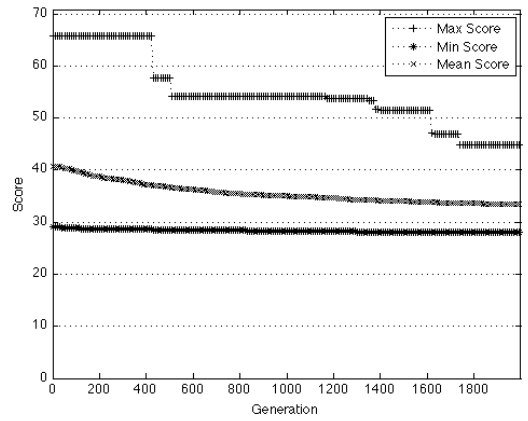

Figure 14: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the German dataset using fitness Equation 2.

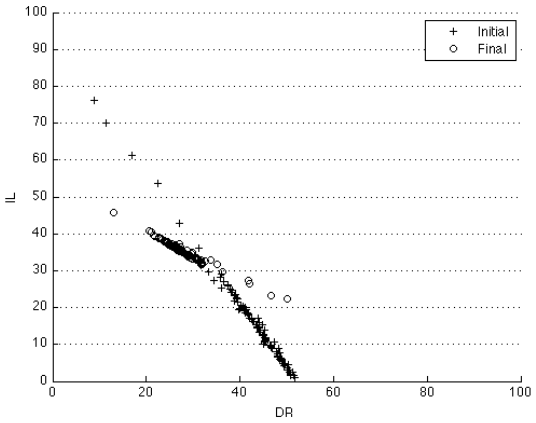

Figure 15: Dispersion plot of initial and final population information loss and disclosure risk for the Flare dataset using fitness Equation 2.

(0.44% of improvement).

After the first two experiments, we can see an interesting fact. We have seen that, in this second experiment, in all the cases the final population is grouped around the pairs of values (IL,DR) with more balanced values than the ones in the first experiment. However, this is achieved in a different way in the four datasets. An analysis of the total number of valid categories in the attributes show that the larger the number of categories, the better the equilibrium of values in both measures. Note that few categories supply a small number of possible different registers. Then, altering some categories increase one of the measures (information loss or disclosure risk) quite abruptly and reduce the other one, and this makes difficult to find an equilibrium between both values.

In addition, we have seen that, using maximum in the fitness

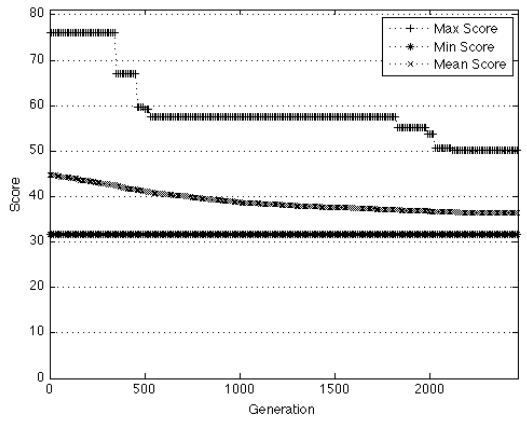

Figure 16: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the Flare dataset using fitness Equation 2.

function (Equation 2) performs better in the optimization than using mean in the fitness function (Equation 1) because the final information loss and disclosure risk measures are more balanced.

It also should be noticed that in all our experiments we obtained an average computation time of $120.34s$ for each entire generation with mutation operation, and $242.48s$ for each entire generation with crossover operation. However, most of the time is consumed by the fitness function ($120.32s$ in mutation generation and $242.46s$ in crossover generation) and a very small amount of time is consumed by the rest of each generation ($0.02s$ in both cases).

3.3 Third experiment: testing the robustness

Finally, to conclude our study, we applied to the Flare dataset our approach using Equation 2 (the maximum value of the two measures is taken as the score value) but in this case not including in the population the best 5% and 10% individuals in terms of the fitness function score. This experiment assesses the robustness of our method trying to achieve the best solutions starting from worse solutions.

After several generations we could see that initial and final populations follow the same behavior than in the case with the entire population. In addition, it can be seen that, compared with Figure 15 , the initial population has a *hole* in the region with more balanced pairs of (IL,DR) which where the ones removed from the population (Figures 17 and 18).

However, looking at the evolution of max, mean and min scores in Figures 19 and 20 we see that we almost reached the best min score obtained without removing these solutions. In the case of removing the 5% of the best initial protections we reached a minimum score of 32.96 what represents a difference of 1.33 points from the minimum value obtained using the entire population, and in the case of removing the 10% of the best initial protections we reached a minimum score of 32.71 what represents a difference of 1.08 points.

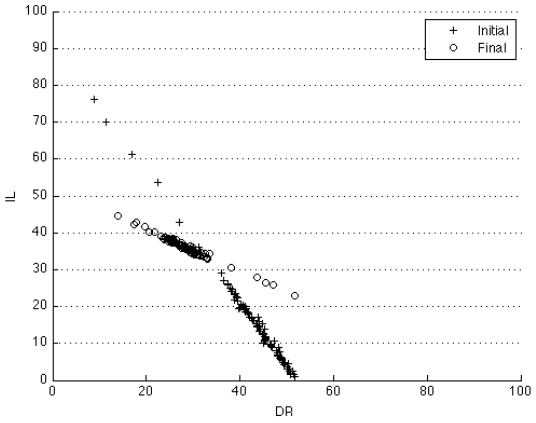

Figure 17: Dispersion plot of initial and final population information loss and disclosure risk for the Flare dataset using fitness Equation 2 without the 5% best initial individuals.

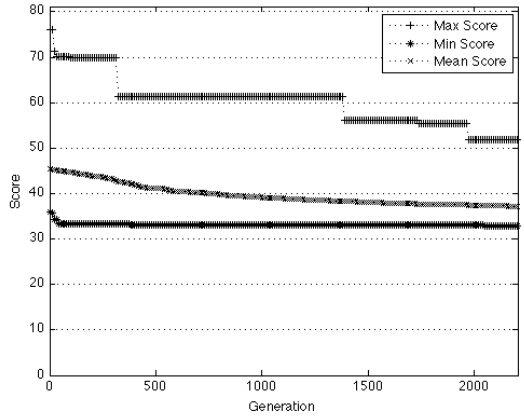

Figure 19: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the Flare dataset fitness Equation 2 without the best 5% initial individuals.

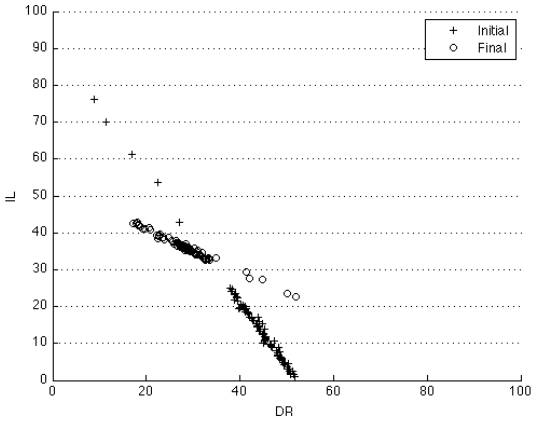

Figure 18: Dispersion plot of initial and final population information loss and disclosure risk for the Flare dataset using fitness Equation 2 without the 10% best initial individuals.

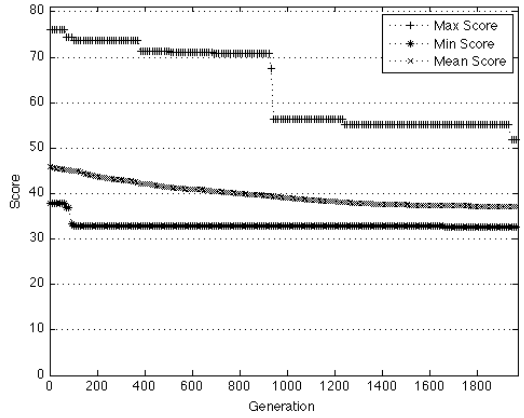

Figure 20: Evolution of the information loss and disclosure risk during the execution of the evolutionary algorithm for the Flare dataset fitness Equation 2 without the best 10% initial individuals.

It should be noticed that the fact of having better result in the case of removing the 10% of the best individuals than in the case of removing just the 5% is produced because of the stochasticity of evolutionary algorithms.

So, looking at this behavior we can assess that our evolutionary approach is robust enough to achieve good protections even when the best ones are missing.

4. CONCLUSIONS

In this paper we proposed an evolutionary algorithm to seek new and enhanced protections for categorical datasets. We presented experimental results using four different real datasets having a good optimization of most of the protections. We also realized that the difficulty of optimizing certain dataset is related to the number of different categories that are valid for the attributes to protect, that is, the more different categories available, the better optimization.

Furthermore, we demonstrated that the use of the mean value of information loss and disclosure risk measures as a fitness function score does not always work well in the case of categorical data, and we proposed a fitness function score based on taking the maximum value of the two measures. Finally, we also demonstrated that our approach is robust against the absence of the best score protections in the population so, it is possible to reach the best score individuals that are missing.

In addition, the advantage of using evolutionary algorithms is that they can be easily adapted to other fitness functions. This is important for our approach because, as it is based on an evolutionary algorithm, it can be adapted to possible new measures of information loss and disclosure risk by just providing a different fitness evaluation function.

The major drawback of our approach is the cost in time

for the computation of the current disclosure risk (DR) and information loss (IL) measures to evaluate the individuals. This issue can be explored as a future work together with some other ways to aggregate them in order to help the algorithm to optimize faster.

5. REFERENCES

[1] Samarati, P., 2001. Protecting respondents' identities in microdata release. IEEE Transactions on Knowledge and Data Engineering 13(6):1010-1027.

[2] Domingo-Ferrer, J., Torra, V., 2001. A quantitative comparison of disclosure control methods for microdata. In Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (ed.), Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies, Elsevier, pp. 111-133.

[3] Fienberg, S., 1994. Conflict between the needs for access to statistical information and demands for condifentiality. Journal of Official Statistics 10, 115-132.

[4] Gouweleeuw, J., Kooiman, P., Willenborg, L., de Wolf, P., 1998. Post randomization for statistical disclosure control: Theory and implementation. Journal of Official Statistics 14, 463-478.

[5] Frank, A., Asuncion, A., 2010. UCI machine learning repository.

[6] Hundepool, A., Willenborg, L., 1998. Argus: Software from the sdc project, in: Proceedings of Joint UNECE-Eurostat Work Session on Statistical Data Confidentiality, Luxembourg:UNECE-Eurostat. pp. 87-98.

[7] Torra, V., 2004. Microaggregation for categorical variables: A median based approach, in: Domingo-Ferrer, J., Torra, V. (Eds.), Privacy in Statistical Databases, Springer. pp. 162-174.

[8] Torra, V., Domingo-Ferrer, J., 2001. Disclosure control methods and information loss for microdata. In Doyle, P., Lane, J.I., Theeuwes, J.J.M., Zayatz, L.V. (ed.), Confidentiality, disclosure, and data access : Theory and practical applications for statistical agencies, Elsevier, pp. 91-110.

[9] Agrawal, R., Srikant, R. 2000. Privacy preserving data mining. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Dallas, Texas, United States, pp 439-450.

[10] Back, T., Fogel, D.B., Michalewicz, Z., 2000. Evolutionary Computation Vol. 2: Advanced Algorithms and Operations. Institute of Physics Publishing.

[11] Dick, G., 2005. A comparison of localised and global niching methods. In: Proc. of the 17th Annual Colloquium of the Spatial Information Research Centre (SIRC 2005: A Spatio-temporal Workshop), Dunedin, New Zealand, pp 91-101.

[12] Holland, J.H., 1992. Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence. MIT Press.

[13] Mahfoud, S.W., 1992. Crowding and preselection revisited. Tech. Rep. 92004, Illinois Genetic Algorithms Laboratory (IlliGAL), University of Illinois, also in Männer, R., Manderick, B. (ed.) Parallel Problem Solving From Nature, Elsevier, PPSN, 2:27-36.

[14] Moore, R., 1996. Controlled data swapping techniques for masking public use microdata sets. Statistical Research Division Report Series RR 96-04, US Bureau of the Census.

[15] Kooiman, P., Willenborg, L., Gouweleeuw, J., 1998. PRAM: A mehtod for disclosure limitation of microdata. CBS research paper 9705. Available from http://www.cbs.nl/research.

[16] Domingo-Ferrer, J., Torra, V., 2002. Distance-based and probabilistic record linkage for reidentification of records with categorical variables. In Butlletí de l'ÀCIA, vol. 28, pp 243-250, Associació Catalana d'Intel.ligència Artificial.

[17] Nin, J., Herranz, J., Torra, V., 2008. Rethinking rank swapping to decrease disclosure risk. Data & Knowledge Engineering, Elsevier. 64, 346-364.

[18] Marés, J., Torra, V., 2010. PRAM Optimization Using an Evolutionary Algorithm. In: Proc. of Privacy in Statistical Databases 2010, CorfÃž, Greece, pp 97-106, Springer.