T-79.511 Special Course on Cryptology / Privacy-Preserving Data Mining: Comments on "Database Randomization via RRT", 2nd Version

Emilia Oikarinen

October 7, 2003

1 Introduction

This is a review of the survey [3] on data randomization via randomized response techniques based on [1, 2]. The survey describes how the randomized response technique introduced by Warner [4] can be extended to be used in association rule mining and data classification.

2 General Comments

This second version has improved remarkably from the preliminary version and gives a rather understandable presentation on the subject. What is not explicitly stated in the survey is the concept of privacy used in the approaches and that should be emphasized. The overall structure of the survey is clear, first the randomized response in explained in Section 2, a method for association rule mining based on [2] is presented in Section 3 and method for data classification based on [1] is presented in Section 4 and finally the survey is concluded with a brief comparison between the two methods. In Section 3.4 one could still try to emphasize what is the goal in defining the support and support recovery (i.e. associations are discovered by the value of the support which is mentioned not until Section 3.5).

3 More Detailed Comments

2 Randomized Responses. The Unrelated-Question Model could be briefly explained. The Equations (1) and (2) are a bit small, maybe a bigger font could be used (same applies also to Equations (9) and (10) in Section 3 and Equations (13), (15) and (16) in Section 5). When defining $P(A = yes)$ it should read something like "$P(A = yes)$ is the estimated proportion of the "yes" in the actual undisguised data". Also it should be stated what $P(A = no)$ and $P^*(A = no)$ mean.

3.2 Privacy Breaches. In Definition 3.2 it should read $i \in 1, \ldots, N$ instead of $1...N$.

3.3 Cut-and-paste Randomization. In item 1. it should read K_m instead of $K_m j$.

3.4 Support Recovery. In Definition 3.5 terms per-transaction and item-invariant are used, but these have not been explained. It should read t' instead of t'. After Equation (4) it should read "must be integers in $\{0, 1, \ldots, k\}$. Before Equation (6) there should read something like "and solve $\overrightarrow{s}$ from the Equation (5)". When defining $q[l \rightarrow l']$ it should read $Q_{(ll')}$, to make the matrix notation consistent with the one used right after Equation (5). I'm also wondering what is $\delta_{l=k}$ in Equation (9).

The survey does not discuss at all, how the privacy breaches can be limided. It would be nice to consider this matter on a informal level (i.e. how it can be achieved on a general level without actual mathematics) just the get the idea.

3.5 Discovering Associations. In item (b) recovering of sigmas is mentioned, but sigma (i.e. supposedly cumulative support) has not been defined. In item (c) σ appears, but it is not defined at all.

3.6 Experiments of Association Discovery (and also 4.4 Classification Experiments). At this level these Sections are a bit confusing. There is not enough information to get a good understanding of the experiments but there it also too much information for just an overview. A suggestion would be to evolve these sections to more general and overview-like level and focus on describing results (i.e. what do the experiments show about the feasibility of the methods), since the details of the experiments can be found from the original articles.

4 Classification of Disguise Data. It should be mentioned, that the technique presented by Du et al. applies to binary data.

4.1 Multivariate Randomized Data. In the second paragraph it should read "for sets that" instead of "for questions that" and in third paragraph "is based" instead of "is base".

4.2 Modified ID3 Algorithm. In the second sentence of the first paragraph the article before $P(E)$ is gratuitous. In the second sentence of the second paragraph the verb should be in plural (are). After Equation (13) there should be "and" between S_v and $|S|$ instead of comma. In addition to estimates for $|S|$, $Entropy(S)$ and $Entropy(S_v)$ also estimate for $|S_v|$ needs to be calculated to solve Equation (13). In the last sentence of this section it should read "get" instead of "getting".

4.3 Accuracy Score. After Equations (15) and (16), the testing data set should be U instead of U. Also no new paragraph should begin after $\overline{U}$. There is a full stop missing at the end of the last sentence in this section.

4.4 Classification Experiments. No space before 6 in "6 continues". In the first sentence of the fourth paragraph it would be better to say something like "θ ranges between". At the end of the paragraph privacy level is mentioned, but it is not defined.

5 Conclusions. In the third sentence of third paragraph it should read "do" instead of "does".

References

[1] W. Du and Z. Zhan, Using Randomized Response Techniques for Privacy-Preserving Data Mining. In *Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, August 2003.

[2] A. Evfimievski, R. Srikant, R. Agrawal and J. Gehrke. Privacy Preserving Mining of Association Rules. In *Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, July 2002.

[3] I. Niemi. Database Randomization via RRT. Survey for the Seminar T-79.514: Special Course on Cryptology/Privacy Preserving Data Mining at Helsinki University of Technology, October 2003.

[4] S.L. Warner. Randomized Response: a Survey Technique for Eliminating Evasive Answer Bias. *The American Statistical Association*, 60(309), pp. 63–69, March 1965.