

# Privacy Preserving Classification for Continuous and Nominal Attributes

Piotr Andruszkiewicz

Warsaw University of Technology, Warsaw, Poland

## Abstract

As a result of advances in technology, large amounts of data can be collected and stored. Significant development of the Internet and easier access to it have contributed to collecting large amounts of information about users' characteristics. Along with these changes, concerns about privacy of data have emerged. Several methods of preserving privacy classification have been proposed in literature. None of them use simultaneously continuous and nominal attributes distorted with the randomization-based technique.

To make privacy preserving classification and the randomization-based technique useful in the real situations, we propose new solutions which allow data miners to use different types of data (continuous and nominal attributes at the same time). Effectiveness of the new solutions has been tested and presented in this paper.

**Keywords:** data mining, privacy preserving, probability distribution reconstruction, continuous and nominal attributes, classification, decision tree

## 1 Introduction

Large amounts of data have been collected recently and data mining algorithms are used to discover knowledge hidden in this data. Considering this situation users are afraid of revealing sensitive values, what leads to inaccuracy of the data and the models. To encourage people to provide information, even about sensitive values, privacy preserving techniques have been proposed.

One of the areas where privacy preserving has been widely studied is classification and several methods have been proposed in literature for centralized data which was distorted using randomization-based techniques. None of them could be used for continuous and nominal attributes simultaneously, what means that these methods have limited application in the real world.

The new proposals presented in this paper allow data miners to build decision tree over data modified with randomization-based technique containing both continuous and nominal attributes.

### 1.1 Related Work

Privacy Preserving Data Mining in classification has been extensively discussed recently (Agrawal and Srikant, 2000) (Lindell and Pinkas, 2000) (Agrawal and Aggarwal, 2001) (Du and Zhan, 2003) (Yang *et al.*, 2005) (Zhang *et al.*, 2005) (Xiong *et al.*, 2007).

Papers (Lindell and Pinkas, 2000) and (Yang *et al.*, 2005) represent cryptographic approach to Privacy Preserving. We use different approach – randomization-based technique.

Privacy preserving for individual values in distributed data is considered in (Zhang *et al.*, 2005) and (Xiong *et al.*, 2007). In these works databases are distributed across a number of sites and each site only willing to share mining process results, but does not want to reveal the source data. Techniques for distributed database require a corresponding part of the true database at each site. Our approach is complementary, because it collects only modified tuples, which are distorted at the client machine.

Agrawal and Srikant (2000) proposed how to build decision tree over disturbed data with randomization-based technique. Presented algorithm uses only continuous attributes.

Paper (Agrawal and Aggarwal, 2001) extends solution proposed by Agrawal and Srikant, but it does not take into account nominal attributes either.

Multivariate Randomized Response technique was presented in (Du and Zhan, 2003). It allows creating decision tree only for nominal attributes.

The proposition showed in this paper differs from those above, because it enables data miner to classify perturbed data containing continuous and nominal attributes modified using randomization-based techniques to preserve privacy on individual level. This approach creates decision tree to classify the data.

### 1.2 Contributions of This Paper

Proposed solution makes Privacy Preserving Classification viable in real situations, because it takes into account continuous and nominal attributes simultaneously.

### 1.3 Organization of This Paper

The remainder of this paper is organized as follows: In Section 2, we present our new algorithm for building decision tree over distorted database containing continuous and nominal attributes. The experimental results are highlighted in Section 3. Finally, in Section 4, we summarize the conclusions of our study and outline future avenues to explore.

## 2 New Proposals

To build decision tree over the data containing nominal and continuous attributes distorted with the randomization-based technique we need to apply algorithms shown below. The remaining problems should be solved in the way proposed in (Agrawal and Srikant, 2000) and (Agrawal and Aggarwal, 2001).

## 2.1 Modification of Probability Distribution Reconstruction Algorithm

Algorithm proposed by Agrawal and Srikant (2000), we will call it AS, was designed to reconstruct probability distribution of continuous attributes. Its extension called EM and presented in (Agrawal and Aggarwal, 2001) uses continuous attributes also.

To reconstruct probability distribution of nominal attribute we modified algorithms mentioned above and called it EM/AS (Algorithm 1), because modifications of both algorithms (AS and EM) give the same result.

The algorithm solves the problem: we have nominal attribute  $X$  with the values  $v_1, v_2, v_3, \dots, v_k$ . Value for each sample is modified according to probability  $Pr(v_p \rightarrow v_s)$  (probability that value  $v_p$  will be changed to value  $v_s$ ) and we want to reconstruct original probability distribution of attribute  $X$ .

---

**Algorithm 1** EM/AS – nominal attribute probability distribution reconstruction algorithm

---

```

 $Pr(X = v_p)^0 := \frac{1}{k}, p = 1, \dots, k$ 
 $j := 0$  //iteration number
repeat
   $Pr(X = v_p)^{j+1} = \frac{1}{n} \sum_{s=1}^n \frac{Pr(v_p \rightarrow v_s) Pr^j(X=v_p)}{\sum_{t=1}^k Pr(v_t \rightarrow v_s) Pr^j(X=v_t)}$ 
   $j := j + 1$ 
until(stopping criterion met)

```

---

Stopping criterion is the same as for AS and EM algorithm (we stop when the difference between successive estimates of the original probability distribution becomes small – 1% of the threshold of the  $\chi^2$  test).

## 2.2 Assigning Reconstructed Values to Samples for Nominal Attributes

Having probability distribution reconstructed, we can assign reconstructed values to samples, what allows us to choose the best test for a tree node using *gini index*<sup>1</sup>. Assigned values should not be treated as estimates of original values.

The problem to be solved is as follows:

We have modified values of nominal attribute, so we have probability distribution of modified attribute (i.e.  $P(Z = v_i)$ ,  $i = 1..k$ ), the number of all samples  $n$  and reconstructed probability distribution ( $P(X = v_i)$ ,  $i = 1..k$ ). We want to assign samples to reconstructed values taking into account reconstructed probability distribution. To solve this problem we count the number of distorted samples separately for each value of attribute ( $n_Z(v_i)$ ) and estimate the number of original samples ( $n_X(v_i) = P(X = v_i)n$ ).

Then we calculate the difference between  $n_Z(v_i)$  and  $n_X(v_i)$  and call it  $\delta(v_i)$ . For  $\delta(v_i) > 0$  we have too many samples.

We look for samples corresponding to positive  $\delta(v_i)$  and assign them the value for which we have negative  $\delta(v_j)$ . We update values of proper  $\delta(v_i)$  and continue process until all values of  $\delta(v_i)$  are zero.

---

<sup>1</sup>For details about *gini index* see (Loh and Shih, 1997).

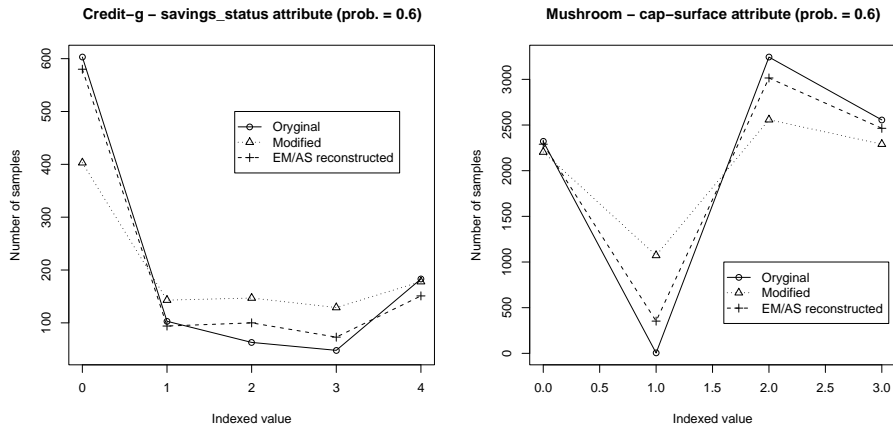


FIGURE 1: Probability distribution reconstruction for nominal attributes with probability of retaining original value equal to 0.6.

Having completed the process, we have samples with the reconstructed values assigned according to original (reconstructed) probability distribution.

In case of reconstructing probability distribution for each class, we perform this process for every single class separately.

Moreover, having reconstructed probability distributions divided into classes, we can choose the best test without assigning the values.

When the probability of retaining the original values is less than 0.5, we look for samples starting from those which belong to the opposite group (e.g. when we have sample which does not meet the test, then we start from samples which meet the test).

Presented solution allows data miner to combine algorithms proposed in this paper with those for continuous<sup>2</sup> attributes and mine databases containing both nominal and continuous attributes.

### 3 Experiments

This section presents the results of the experiments conducted with algorithm for assigning reconstructed values to samples for nominal attributes combined with EMAS algorithm.

#### 3.1 Probability Distribution Reconstruction for Nominal Attributes

Figure 1 shows original, distorted and reconstructed probability distribution of two nominal attributes: `savings_status` (set `credit-g`) and `cap-surface` (set `mushroom`)<sup>3</sup>.

<sup>2</sup>Continuous attributes are modified with additive perturbation technique (Agrawal and Srikant, 2000).

<sup>3</sup>All sets used in tests can be downloaded from UCI Machine Learning Repository (<http://www.datalab.uci.edu/>).

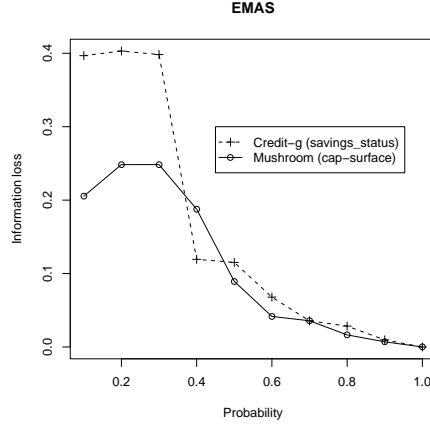


FIGURE 2: Information loss caused by incorporating privacy and reconstructing probability distribution using EMAS algorithm.

The probability of retaining the original value is equal to 0.6.

The reconstructed probability distributions are close to the original distributions. It means that we do not lose too much information.

The lack of precision in reconstructing is called information loss. The information loss is defined as follows (Agrawal and Aggarwal, 2001):

**Definition 3.1** Information loss  $\mathcal{I}(f_X, \hat{f}_X)$  equals half the expected value of  $L_1$  norm between the original probability distribution  $f_X$  and its estimate  $\hat{f}_X$ .

$$\mathcal{I}(f_X, \hat{f}_X) = \frac{1}{2} E[\int_{\Omega_X} |f_X - \hat{f}_X|]$$

Information loss  $\mathcal{I}(f_X, \hat{f}_X)$  lies between 0 and 1. 0 means the perfect reconstruction and 1 implies that there is no overlap between original distribution and its estimate.

The information loss caused by incorporating privacy for nominal attribute is shown in Figure 2. For probability equal to 1 the information loss is 0, because there was no distortion. The less probability of retaining the original value, the higher information loss. In general, the privacy causes information loss.

The accuracy of classification (the percentage of correct classified samples) for Dna and Soybean-large sets is shown in the Figure 3. Both sets contain only nominal attributes.

In the experiments we have several types of reconstruction. *Global* means that we reconstruct probability distributions only in the root of a tree. In *By class* we reconstruct separately for each class but only in the root node. For *Local* reconstruction is performed in every node divided into classes. *Local all* – reconstruction in every node without dividing into classes. *Random* means without reconstruction.

For Soybean-large set we have higher accuracy for higher probability of retaining original value. For Dna set we can observe the inverted bell shape (for *Local*

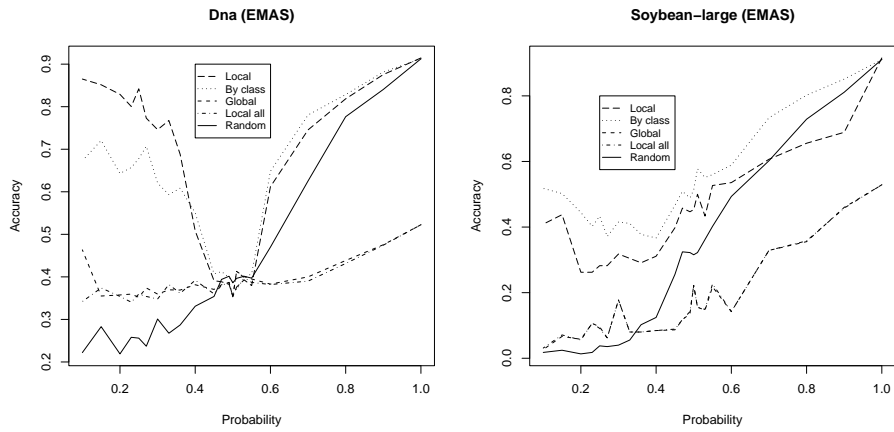


FIGURE 3: Accuracy of classification for Dna and Soybean-large sets containing only nominal attributes classified with EM/AS algorithm.

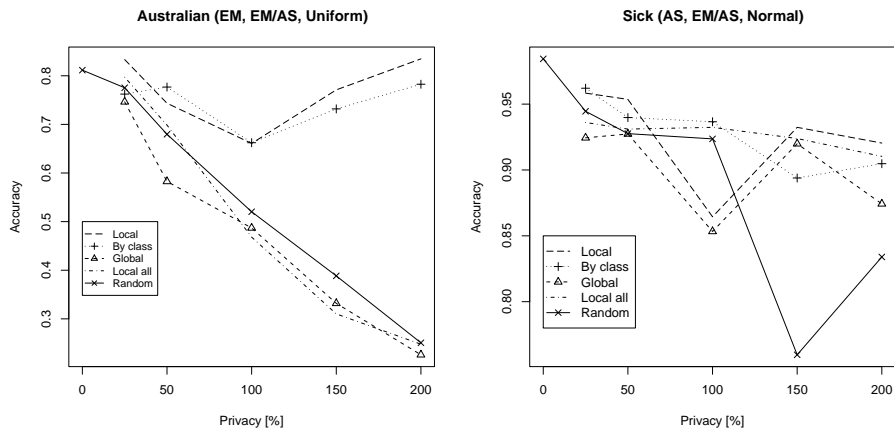


FIGURE 4: Accuracy of classification for Australian and Sick sets containing nominal and continuous attributes (Australian set classified with EM and EM/AS algorithms and uniform distorting distribution, Sick set with AS, EM/AS algorithms and normal distorting probability distribution).

and *By class* reconstruction), because this set contains binary attributes.

Figure 4 shows the accuracy of classification for Australian and Sick sets (both contain nominal and continuous attributes).

Like Agrawal and Srikant (2000) stated *Local* and *By class* seem to be the best types of reconstruction. For 150%-200% privacy level the accuracy (*Local* and *By class*) for Australian set is as high as without preserving privacy and for Sick set is significantly better than without reconstruction. Results for normal/uniform distorting probability distribution, respectively, were similar.

## 4 Conclusions and Future Work

We investigated the problem of building decision tree over perturbed data containing nominal and continuous attributes simultaneously. We focused on the situation where we have centralized data distorted using randomization-based technique.

Effectiveness of the new solution has been tested on synthetic and real databases. The results of these experiments show that proposed algorithms can be used in real situations.

In future works, we plan to investigate the possibility of extension of our results to preserve privacy for target (class) attribute.

## References

- Dakshi AGRAWAL and Charu C. AGGARWAL (2001), On the design and quantification of privacy preserving data mining algorithms, in *PODS '01: Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 247–255, ISBN 1-58113-361-8.
- Rakesh AGRAWAL and Ramakrishnan SRIKANT (2000), Privacy-preserving data mining, in *Proc. of the ACM SIGMOD Conference on Management of Data*, pp. 439–450, ACM Press.
- Wenliang DU and Zhijun ZHAN (2003), Using randomized response techniques for privacy-preserving data mining., in Lise GETOOR, Ted E. SENATOR, Pedro DOMINGOS, and Christos FALOUTSOS, editors, *KDD*, pp. 505–510, ACM, ISBN 1-58113-737-0.
- Yehuda LINDELL and Benny PINKAS (2000), Privacy Preserving Data Mining, in Mihir BELLARE, editor, *CRYPTO*, volume 1880 of *Lecture Notes in Computer Science*, pp. 36–54, Springer, ISBN 3-540-67907-3.
- W. Y. LOH and Y. S. SHIH (1997), Split Selection Methods for Classification Trees, *Statistica Sinica*, 7(4):815–840.
- Li XIONG, Subramanyam CHITTI, and Ling LIU (2007), Mining multiple private databases using a kNN classifier, in *SAC '07: Proceedings of the 2007 ACM symposium on Applied computing*, pp. 435–440, ISBN 1-59593-480-4.
- Zhiqiang YANG, Sheng ZHONG, and Rebecca N. WRIGHT (2005), Privacy-Preserving Classification of Customer Data without Loss of Accuracy, in *SDM*.
- Nan ZHANG, Shengquan WANG, and Wei ZHAO (2005), A new scheme on privacy-preserving data classification, in Robert GROSSMAN, Roberto BAYARDO, and Kristin P. BENNETT, editors, *KDD*, pp. 374–383, ACM, ISBN 1-59593-135-X.

