A Classification based Dependent Approach for Suppressing Data

Vamshi Batchu
Hindustan University
Chennai, India

D.John Aravindhar
Hindustan University
Chennai, India

J.Thangakumar
Hindustan University
Chennai, India

.M.Roberts Masillamani
Hindustan University
Chennai,India

ABSTRACT

Data mining plays an important role in internet with the computer technology this makes easy to collect the information from the related data sets. The different methods used in this paper are decision tree algorithm, the decision tree algorithm used hears is to classify the data elements by considering a set of constraints, we consider this method to suppress the data by doing so we can secure the data. We extend our work on micro data suppression (1) to prevent not only probabilistic but also decision tree classification based inference, and (2) to handle not only single but also multiple confidential data value suppression to reduce the side-effects. The paper aims to enhance the Data classification and Data Generalization. It shows that how the data is secured using 'Generalization' and moreover. It provides efficiency in Data Generalization and discusses some of the major challenges for what kind of data to be suppressed. We consider the following privacy problem: a data holder wants to release a version of data for building classification models, but wants to protect against linking the released data to an external source for inferring sensitive information. The generalized data remains useful to classification but becomes difficult to link to other sources. The generalization space is specified by a hierarchical structure of generalizations. A key is identifying the best generalization to climb up the hierarchy at each iteration. Enumerating all candidate generalizations is impractical.

***Key words**: Data classification, Data security, Data generalization, Data mining.*

1. INTRODUCTION

In tandem with the advances in networking and storage technologies, the private sector as well as the Public sector has increased their efforts to gather, and manipulate information on a large scale. Non governmental organization collects information about there customer or members for many reasons including better customer relationship management and high level decision making. This pervasive data-harvesting effort coupled with the increasing need to share the data with other institutions or with public raised concerns about privacy is the ability of an individual to prevent information about himself becoming known to other people with out his approval [1]. More specifically it is the right of individuals to have the control over the data they provide. This includes controlling, (1) How the data are going to be used? (2) Who is going to use it? (3) For what purpose?

2. EXISTING SYSTEM

In the Existing System, the basic idea of the design is to collect the data from the user and classify the data using classification algorithms like decision tree classification.

For classification we use different techniques like ID3, maximum impact data attributes, next best guess, but we found that this method does not work well for securing the data. Some of the Problem in using ID3 algorithms depends on the number of attributes, ensuring that the success rate of these algorithms will always be higher than the other algorithms if the number of attributes is higher than the number of transactions. Also depend both on the number of attributes and the number of transactions.

A.DECISION TREE

A decision tree is a flow chart-like tree structure, where each internal node (non leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label. The top most nodes in a tree are the root node [2].

Example 1

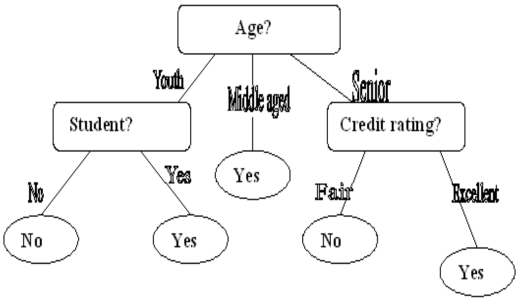


Figure **1: Decision tree**

The figure 1 is a decision tree for the concept bugs-computer, indicating whether a customer at 'All Electronics' is likely to purchase a computer. Each internal (no leaf) node represents a test on an attribute. Each leaf node represents a class (either buys-computer= yes or buys-computer=no)

B."How is decision trees used for classification?"

Given a tuple, X, for which the associated class; able is unknown; the attribute values of the tuple are tested against the decision tree. A path is traced from the root to a leaf node, which holds the class prediction for that tuple. Decision tree can easily be converted to classification rules [2].

C. Generate-decision-tree

Aim: Generate a decision tree from the training tuples of data partition D.

Input:

- Data partition, D, which is a set of training tuples and their associated class labels;
- Attribute-list, the set of candidate attributes.
- Attribute-Selection-method, a procedure to determine the splitting criterion that "beat" partitions the data tuples into individual classes.

This criterion consists of a splitting-attribute and, possibly, either a split point or splitting subset.

Out put: A decision tree

Method

```
1.  Create a node N;
2.      if tuples in D are all of the same class, C then
3.      return N as a leaf node labeled with the majority class C;
4.      if attribute-list is empty then
5.  Return N as a leaf node labeled with the majority class in D; // majority voting.
6.  Apply Attribute-selection method (D, attribute-list) to find the "best" splitting-criterion;
7.  lable node N with splitting-criterion;
8.  If splitting-attribute is allowed then // not restricted to binary trees.
9.  Attribute-list←attribute-list—splitting—attribute; // remove splitting-attribute.
10. For each outcome j of splitting-criterion// partition the tuples and grow sub trees for each partition.
11. let Dj be the set of data tuples in D satisfying outcome j;// a partition
12. if Dj is empty then
13. attach a leaf labeled with the majority class in D to node N;
14. else attach the node returned by Generate-decision-tree (Dj, attribute-list) to node N; end for
15. return N;
```

D. Bayesian classification

"What are Bayesian Classifiers?"

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on "Bayes' theorem".

Baye's Theorem

Let X be the data tuple, is considered "evidence". It is described by measurements made on a set of n attributes. Let H be some hypothesis, such as the data tuple X belongs to a specified class C. for classification problem we have to determine P (H/X), the probability that the hypothesis H holds given the "evidence" or observed data tuple X [2].

"How are these probabilities estimated?"

P (H), P (X/H), and P (X) may be estimated from the given data. Bayes theorem provides a way of calculating the posterior probability, P (H/X) from P (H), P (X/H) and P (X).

$$P (H/X) = P (X/H) * P (H)/P(X)$$

E. Data Suppression

The most common method of preventing the identification of specific individuals in tabular data is through cell suppression. This means not providing counts in individual cells where doing so would potentially allow identification of a specific person. Cell suppression can also be done by combining cells from different small groups to create larger groupings that reduce the risk of identifying individuals While there are also more sophisticated data perturbation methods that use statistical noise to mask sensitive information, these are generally more suitable for use with economic or financial data than with public health data. This appendix reviews the basic methods, issues, strengths, and vulnerabilities of cell suppression. In possible statistical unreliability of estimates that are based on small numbers [4].

F. Suppression Criteria

Suppression rules are typically based on a predetermined criterion for the number of diagnosed cases and/or the number of births in the population or subpopulation from which the cases were identified. These numbers may also be thought of as the numerator and the denominator, respectively, of a prevalence estimate. In practice, the rules used vary from relatively liberal to very conservative [5].

Having made the decision to suppress, the question becomes what and how to suppress. The solution that provides the greatest protection of privacy is to suppress an entire table whenever a single cell presents a threat, whereas the solution that provides the least protection is to suppress a single offending cell or only those cells deemed sensitive.

3. PROPOSED SYSTEM

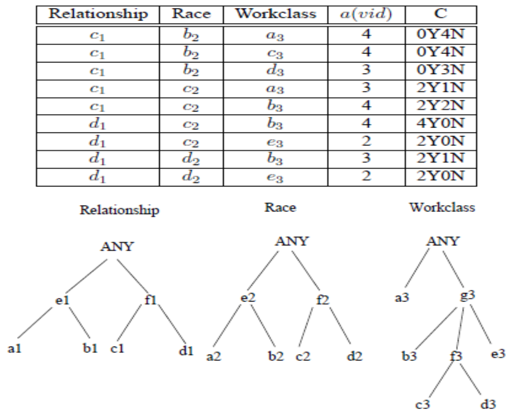
In the proposed system the paper mainly concentrates on the generalization concepts. The idea is simple but novel: we explore the data generalization concept from data mining as a way to hide detailed information, rather than discover trends and patterns. Once the data is masked, standard data mining techniques can be applied without modification. Our work demonstrated another positive use of data mining technology: not only can it discover useful patterns, but also mask private information.

A. Anonymity:

The virtual identifier, denoted VID, is the set of attributes shared by R and E. a (vid) denotes the number of records in R with the value vid on VID. The anonymity of VID, denoted A (VID), is the minimum a(vid) for any value vid on VID. If a (vid) = A (VID), vid is called an anonymity vid. R satisfies the anonymity requirement < VID; K > if A (VID) ¸ K, where K is specified by the data holder. We transform R to satisfy the anonymity requirement by generalizing specific values on VID into less specific but semantically consistent values. The generalization increases the probability of having a given value on VID by chance, therefore, decreases the probability that a linking through this value represents a real life fact. The generalization space is specified through a taxonomical hierarchy per attribute in VID, provided by either the data holder or the data recipient. A hierarchy is a tree with leaf nodes representing domain values and parent nodes representing less specific values. R is generalized by a sequence of generalizations, where each generalization replaces all child values c with their parent value p in a hierarchy. Before a value c is generalized, all values below c should be generalized to c first.

Generalization:

A generalization, written {c} → p, replaces all child values {c} with the parent value p. A generalization is valid if all values below c have been generalized to c. A vid is generalized by {c}→ p if the vid contains some value in {c}.



Relationship	Race	Workclass	$a(vid)$	C
c_1	b_2	a_3	4	0Y4N
c_1	b_2	c_3	4	0Y4N
c_1	b_2	d_3	3	0Y3N
c_1	c_2	a_3	3	2Y1N
c_1	c_2	b_3	4	2Y2N
d_1	c_2	b_3	4	4Y0N
d_1	c_2	e_3	2	2Y0N
d_1	d_2	b_3	3	2Y1N
d_1	d_2	e_3	2	2Y0N

Fig2: Data and hierarchies for V ID

2., Anonymity for Classification:

Given a relation R, an anonymity requirement <VID,K>, and a hierarchy for each attribute in VID, generalize R, by a sequence of generalizations, to satisfy the requirement and contain as much information as possible for classification. The anonymity requirement can be satisfied in more than one way of generalizing R, and some lose more information than others with regard to classification. One question is how to select a sequence of generalizations so that information loss is minimized. Another question is how to find this sequence of generalizations efficiently for a large data set.

4 ANALYSIS

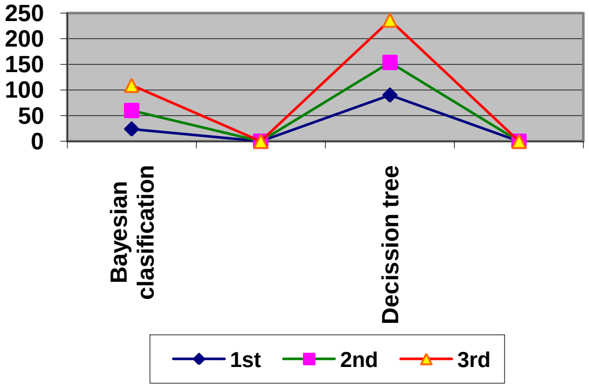


Figure 3: Comparison of classification

5 CONCLUSION

In final classification and Generalization will be analyzed.
The paper mainly concentrates on the issues like:
Suppressing confidential data values against other classification algorithms, e.g., logistic regression

- Suppressing multiple confidential data values at a time (generic version having no constraints),
- Developing a generic suppression technique, independent from individual classification methods, based on information theory,
- Using generalization as a fine grained method, and Suppressing evolving (i.e., continuously updated) micro data.

First objective is to evaluate the quality of generalized data for classification, compared to that of the unmodified data.
Second objective is to evaluate the scalability of the proposed algorithm and generate the generalized report on the data.

REFERENCES

[1] Klein RJ, Proctor SE, Bouderault MA, Turczyn KM. Healthy People 2010 criteria for data suppression. Healthy People 2010 Statistical Notes. No.24. Hyattsville, MD: National Center for Health Statistics; pp.

(2002).

[2] "Data mining: Concepts and Techniques", Jiawei Han, Macheline Kamber, Morgan Kaufmann Publishers, chapter-6, page no 358.pp. (2005)

[3] Aggarwal, C.: On k-anonymity and the curse of dimensionality. In: Proceedings of the 31st VLDB Conference (2005).

[4] Doyle P, Lane JI, Theeuwes JM, Zayatz LM, eds. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies. Amsterdam, Netherlands: Elsevier Science pp.185–213 (2001).

[5] Ayca Azgin Hintoglu, Yucel Saygın, "Suppressing microdata to prevent classification based inference", ACM .pp. (2009).