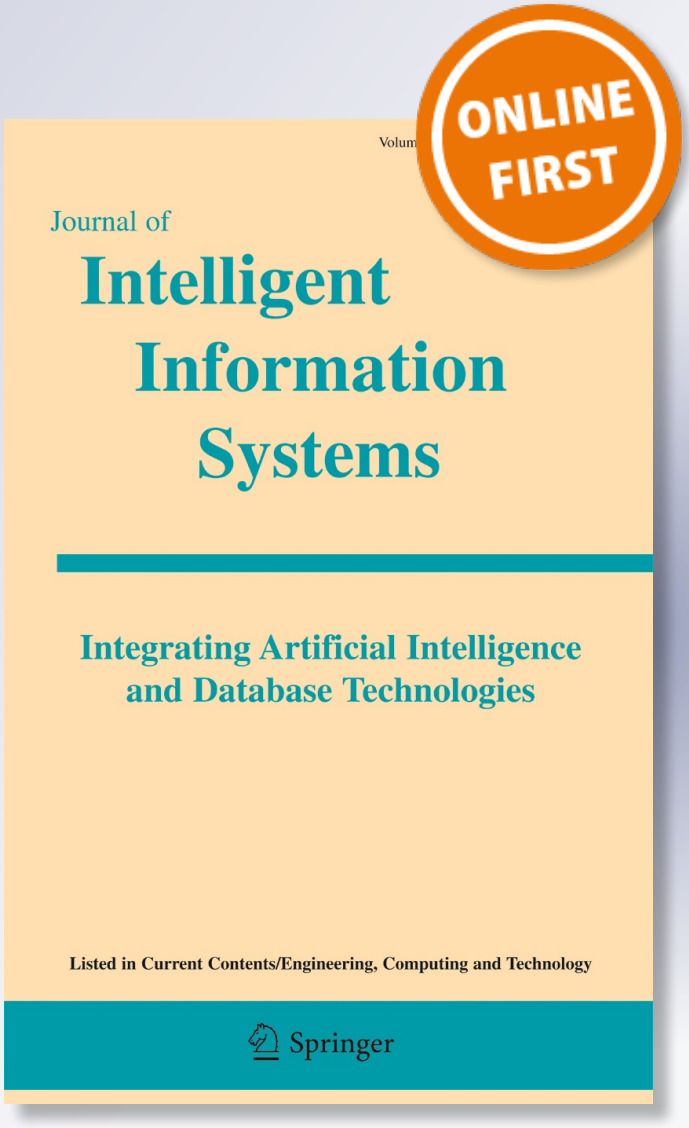
A distributed decision support algorithm that preserves personal privacy

George Mathew & Zoran Obradovic

Journal of Intelligent Information Systems
Integrating Artificial Intelligence and Database Technologies

ISSN 0925-9902

J Intell Inf Syst
DOI 10.1007/s10844-014-0331-6



Springer

J Intell Inf Syst
DOI 10.1007/s10844-014-0331-6

A distributed decision support algorithm that preserves personal privacy

George Mathew · Zoran Obradovic

Received: 10 March 2013 / Revised: 31 July 2014 / Accepted: 4 August 2014

Abstract Assuring confidentiality of personal information and preserving privacy are vital when data is harvested from multiple institutions for business decision-making. An algorithm that builds knowledge using statistics based on subject data from distributed sites that satisfy specified selection criteria is presented here. The algorithm maintains complete fidelity of information structures in the distributed data compared to the centralized equivalent. Heterogeneous data schemas across sites can be accommodated and thresholds can be set for global minimum saturation for attributes to participate in the prediction model building. Policies for inclusion and exclusion of non-exhaustive attributes among sites are introduced. Unification of attributes is introduced for homogenizing attribute values globally. Results of experiments using data from medical, higher education, and social domains elucidate the value of our algorithm in regulated industries, where shipping raw data outside parent institution is not practical.

Keywords Data privacy · Privacy-preserving framework · Distributed decision support systems

1 Introduction

Due to regulations similar to EU data protection directive (Allaert and Barber 1998) in Europe and HIPAA (Sweeney 2010) as well as FERPA in USA, sharing sensitive personal information between institutions is a difficult proposition. However, gathering intelligence from distinct entities by harvesting local information is valuable to many agencies (Rockwell and Abeles 1998). This could be the result of rare samples being spread across many sites or targeted population being geo-dispersed. In the financial sector, a study may be geared towards understanding the repossession pattern of houses in a specific geographic area. A social agency may be interested in understanding certain characteristics of teen drivers in the country. In line with Evidence Based Medicine, a Clinical Decision Support (CDS) system (Courtright 2001) that can provide suggestive knowledge representations based on data sets with patient

G. Mathew (✉) · Z. Obradovic
Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA
e-mail: George.Mathew@temple.edu

Z. Obradovic
e-mail: Zoran.Obradovic@temple.edu

Published online: 24 September 2014

attributes that are similar to the attributes of the patient in context, is valuable to a medical practitioner (Karthikeyan and Pais 2010). Aggregating similar samples from other distributed (off-site) databases would help in better decision-making (Fu 2001; Park and Kargupta 2003). As an example, for studying the enrollment pattern of freshman women students in engineering, a social scientist would like to obtain information that goes beyond what is seen at the local university and what might be available of a more general student population in universities around the nation. The intelligence being sought is usually seeded by partial information, which is commonly known as the "attributes of interest" (Khoshgoftaar 2005). In the example of freshman women students, the attributes of interest with targeted values can be represented as a vector: <year = 1, sex = 'f', college = 'engineering'>. To support distributed decision-support, we suggest an algorithm that can query participating sites to retrieve statistics about data instances satisfying the query and build a knowledge representation for the global prediction model using aggregation of local statistics.

Dynamically acquiring targeted sample data directly from other sites is difficult due to legal and regulatory implications (Vest and Gamm 2010). Even if two sites agree to collaborate, the data schema in the two sites can be different. It is also important to protect the confidentiality of personally sensitive information in inter-institutional data transactions (Goldstein 2000). Given these obstacles, it is advantageous to model a distributed decision support system that makes use of statistics about the samples from distributed sites based on the attributes of interest, rather than the actual data or microdata from other sites (Chow and Mokbel 2011). Our algorithm for decision support does not require raw data from distributed sites and it does not require identical data schema at participating sites. The algorithm does not introduce any noise and the knowledge representation learned by the distributed mechanism is theoretically provable to be identical to the centralized counterpart, with no loss of fidelity.

The knowledge to be harvested needs a representation. One of the widely used artifact for capturing learned information is a decision tree (Moret 1982). A 'Decision Tree' is a data structure for representing paths of traversals in a decision making process for the class of problems known as classification problems. An example of a classification problem is to categorize applicants for a job as qualified or under-qualified or over-qualified. One of the commonly used decision tree-building algorithm that is fairly easy to interpret is ID3 (Quinlan 1986). ID3 builds the decision tree iteratively by starting at the root node and splitting nodes based on a node-splitting criterion that gives better distribution of the data instances into more focused decision paths. The attribute with the highest gain (Quinlan 1993) or equivalently the smallest impurity (or entropy) is picked for the node split. Typically for a decision tree algorithm, the whole raw data is made use of in one central location. Our Distributed ID3-based Decision Tree (DIDT) algorithm extends this to a distributed model where no raw data is needed at a central location to generate the decision tree. A centralized service called Clearing House (hereinafter referred to as CH) mediates the negotiations between the distributed sites. In our distributed decision tree building, the inter-site (global) data schema (of attributes and classes) is dynamically generated with no prior knowledge of local data schemas. The computational model followed in our distributed prediction model building algorithm involves local information processing and global synthesis (Caragea et al. 2004). The initial query based on the attributes of interest and subsequent queries for decision tree building are processed against the local data sets and the resulting local statistics are globally synthesized. This distributed processing concept is illustrated in Fig. 1.

In Fig. 1, D_1, ….., Dn are the distributed sites, S_1, ….., S_n are the local statistics from each site and $g(S_1, …., S_n)$ represents the global synthesis.

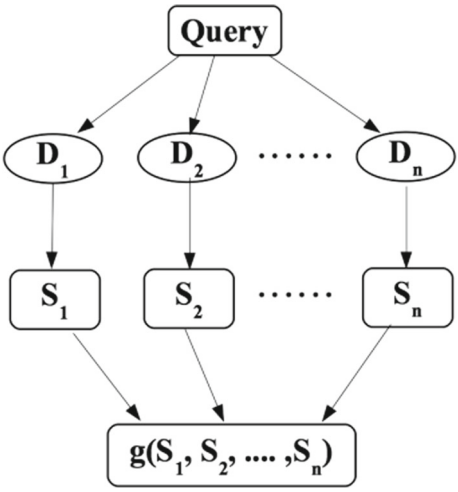


Fig. 1 Distributed decision tree building

2 Related work

Privacy Preserving Distributed Data Mining (PPDM) (Lindell and Pinkas 2000; Xu 2011) has emerged as a field of research interest. PPDM is aimed at mining information from different sources without sacrificing the privacy of the parties involved. Privacy preserving data mining has been studied by researchers in different communities – database community, the statistical disclosure community and the cryptography community (Aggarwal and Yu 2008). The pioneering work on Secure Multiparty Computation (SMC) by Yao (1986) along with other works appear early in the literature (Goldreich 1998; Canetti 1998) and studied in the cryptographic community. PPDM is a form of Secure Multiparty Computation (SMC) (Du and Atallah 2001; Lindell and Pinkas 2009). Preserving privacy of individuals is the essence of Secure Multiparty Computation. A survey of various approaches to SMC also appear in the literature (Vaidya and Clifton 2003b). Other generic cryptographic PPDM techniques also exists (Pinkas 2002). In the statistical disclosure control community, both non-interactive (Adam and Wortman 1989; Brand 2002) and interactive (Dinur and Nissim 2003; Dwork 2006) query modes have been studied. One method of achieving PPDM is to model a distributed decision support system that makes use of statistics about the samples from distributed sites, rather than the actual data from those sites (Chow and Mokbel 2011). The distributed algorithm for decision support introduced in this research does not require raw data from distributed sites and it does not require identical data schema at participating sites. The algorithm does not introduce any noise and the knowledge representation learned by the distributed mechanism is identical to the centralized counterpart, with no loss of fidelity. Data mining algorithms for distributed classification using SVM (Yu et al. 2006) and Logistics Regression (Wu et al. 2012) that preserves privacy appear in the literature. Decision tree is a popular classification model that is easy to interpret and computationally efficient (Cieslak et al. 2012). Caragea et al. (2004) have outlined a theoretical sketch for a distributed decision tree building process. Bar-Or et al. (2005) also made use of the ideas suggested by Caragea et al. to introduce a distributed decision tree induction, viz. DHDT (Distributed Hierarchical Decision Tree). Our DIDT algorithm design was influenced by both of these works. However, both models (Caragea et al. & Bar-Or et al.) require identical relational data schemas (i.e.,

homogeneous data schemas) in each of the participating sites and the schema has to be known to the central broker. DIDT can accommodate non-homogeneous schemas at participating sites and does not require prior knowledge of any data schema. This gives the flexibility for sites to participate fully or participate by not disclosing certain attributes or not participate at will and yet maintain data schema independence locally. Mechanisms similar to differential privacy (Friedman and Schuster 2010) or constraint graphs (Mathew and Obradovic 2011a) can be used as localized constraint filters.

The two data partitioning models for distributed data sets outlined by Caragea et al. (2004) are horizontal & vertical fragmentation. In horizontal data partitioning, all data instances have identical attribute sets and the instances corresponding to a specific value of an attribute will be located in different sites. In vertical partitioning, the attribute set is subdivided and each site holds values for the attributes in the subdivision assigned to it. To get a complete data instance, the sub-tuples are to be combined from different sites. The hybrid model we propose allows for a flexible data schema. A pictorial representation of these models is shown in Fig. 2.

In Fig. 2, columns represent attributes and rows represent instances of data. In the hybrid model, there is no predefined set of attributes for data instances.

Various studies have been done on vertically partitioned data. Giannella et al. (2004) proposed a distributed decision tree-building algorithm for vertically partitioned data. However, the distributed decision tree generated may not be identical to the centralized tree. A privacy-preserving decision tree building over vertically partitioned data was proposed by Vaidya and Clifton (2005). Privacy-preserving k-means clustering (Vaidya and Clifton 2003a) and Privacy-preserving Kth Element Score (Vaidya and Clifton 2009) over vertically partitioned data has also been proposed. Vertical partitioning is usually done on databases to improve performance of transactions (Navathe et al. 1984) and so from a practical use point of view, distributed vertically partitioned databases are implemented across multiple departments within an organization. Clustering (Inan et al. 2006) and Association Rules (Kantarcioglu and Clifton 2004; Kumbhar and Kharat 2012)) on Horizontally Partitioned Data have been studied. From a computational standpoint, privacy-preserving algorithms on vertically partitioned databases may need to restrict the number of participating sites as well as the total size of data involved; while for horizontally partitioned data, only the number of sites needs to be controlled (Kantarcioglu 2008).

The two main techniques for privacy-preserving distributed data mining are secure multi-party computation (SMC) and perturbation (Kantarcioglu et al. 2009). In the case of SMC, the sites cooperate to build the global prediction model without revealing local data. In perturbation, a transformation is performed on the data instances before being used in the model building. There are two general perturbation techniques. One is randomization, in which some noise is added to data before being published. Randomization techniques similar to differential privacy (Dwork 2006) hampers the utility of the model. The second perturbation technique is a mapping of the data instances to some representational data structure with no noise addition. Our algorithm uses such a perturbation technique.

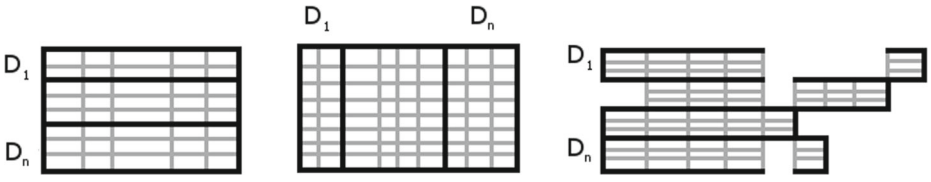


Fig. 2 Data fragmentations: horizontal (*left*), vertical (*middle*), and hybrid (*right*)

Preliminary empirical investigation of our proposed algorithm (Mathew and Obradovic 2011b) using SPECT heart data set (Frank and Asuncion 2010a) was promising. In this paper, we provide empirical validations for the equivalency of the centralized data structure (viz. crosstable) and the aggregation of the localized crosstables; which is the key to our algorithm. A new method of entropy calculation viz. relative entropy is introduced for entertaining attributes that do not span all data instances across the distributes sites, but has coverage span above a given threshold. We define two new policies (union and intersection policies) to be applied on the sets of attributes from participating sites. The concept of 'unification of attributes' to eliminate cross-band effects is also introduced in this paper. The original global schema disclosure to individual sites was refined so as not to divulge unwarranted attributes to participating sites, thus improving security. Each site now gets a conditional global schema (conditional based on the original schema created locally). Experiments using real data from various domains (viz. healthcare, higher-education, and social) are presented. The experiments in this work are done with identical cross-validation sets for our distributed algorithm and the corresponding centralized ID3 algorithm, to demonstrate their equivalence. The experiments in the preliminary study had cross-validation mismatch issues. Our algorithm allows for a flexible localizable constraint enforcement mechanism to be used to assure that local policies related to Personally Identifiable Information and privacy constraints are implemented. In our experiments, attribute-based constraints were used. Attribute-based constraints specify what attributes at particular sites should be blocked from being disclosed. However, this can be replaced by any another localized constraint filter.

3 Methodology

The goal of Distributed ID3-based Decision Tree (DIDT) algorithm is to generate a decision tree from the distributed data instances that would functionally be identical to the decision tree obtained by applying ID3 on the aggregate set of all the distributed data instances. A query mechanism is incorporated to restrict the selection to instances that satisfy the query. There are 2 stages in the DIDT algorithm. In the first stage, a global schema of the distributed data instances relative to a query is generated. The second stage involves gathering statistics from distributed sites based on the global schema and building the next node in the decision tree. In this algorithm, we handle categorical attributes. We also accommodate a complex test from C4.5 in which possible values of an attribute are "allocated to a variable number of groups with one outcome for each group" (Quinlan 1993). Continuous attributes and other add-ons in C4.5 can be accommodated as well with some extensions to our algorithm. A centralized agent called a Clearing House (hereinafter referred to as CH) will broker a query to mine decision-support information from distributed sites. An important assumption is that in each individual site, the data instances will remain the same throughout the decision tree creation process. Another assumption is that all participating sites use agreed upon common vocabulary (Mathew and Obradovic 2010) for attribute names, data types and class names.

The query originated by an end user is passed along to the CH. At the discretion of CH, the query is forwarded to k chosen participating sites. When a site i receives the query, local database is checked for data instances that matches the query. The resulting set of data instances is subjected to local constraint enforcements. The metadata about the resulting set

of attributes and the corresponding set of values are communicated to the CH. The CH synthesizes this information and generates a global schema. For example, assume the metadata from the only 2 participating sites are as follows:

<{sex,income},{{m,f},{med,high}}>

<{sex,income,age},{{m,f},{low}{teen,adult}}>

The global schema can be generated using union or intersection of the attribute sets. When set union is applied to generate global schema, we say that 'union policy' is applied. In this example, the global schema using union policy is:

<{sex,income,age},{{m,f},{low,med,high},{teen,adult}}>

Similarly, the global schema using 'intersection policy' is:

<{sex,income},{{m,f},{low,med,high}}>.

The CH generates conditional global schema for each one of the k sites selected earlier. A conditional global schema for site i is a customized version of the global schema specific to the site and is generated by masking out the attributes and values from the global schema that were not originated by site i. This is done so as to not disclose any attribute information to a site that is not aware of it. At the same time, the conditional global schema gives a uniform template for all local sites to communicate statistics to CH. Referring to the above example, using '?' to mask out a token, the conditional global schema for the first site (using either policies) will be:

<{sex,income},{{m,f},{?,med,high}}>

The CH communicates relevant conditional global schema to each site. Each site generates *crosstable matrices* (Caragea et al.; 2004) for individual attributes in the conditional global schema. For a given attribute u, the $(x,y)^{\text{th}}$ entry of the crosstable matrix represents the number of data instances that exists locally for which the attribute u exists with value x and the instance belongs to class y. Local crosstable matrices generated by the sites are sent to the CH. The CH will aggregate site-specific crosstable matrices to create global crosstable matrices for each attribute. Using global crosstable matrices, the weighted average impurity measure for each attribute is calculated. The attribute that gives minimum impurity is chosen for the next node split. The formal algorithm is given below.

A. Stage 1

01. A query Q by an end user is transmitted to the CH

02. CH, at its discretion, selects k appropriate sites and sends the query Q to the k sites $S_1, \ldots\ldots, S_k$

03. Each site i generates the set L_i of local data instances that satisfies the query Q

04. Let Π_i represent the site-specific constraints and $G_i=\Pi_i(L_i)$

05. Let A_i denote the set of all attributes relative to G_i. For a given attribute $x \varepsilon A_i$, let V^i_x denote the set of values observed in G_i. Let C_i denote the set of all classes that are in G_i. Then, the metadata tuple for the site i takes the form$<A_i, \{V^i_x \mid x \varepsilon A_i\}, C_i, \mathrm{n_i}>$, where $\mathrm{n_i}=|G_i|$. Each site sends the metadata to CH

06. CH creates a global schema by applying union policy on the local schemas using the tuple expression:

$$< \bigcup_{i=1}^{k} A_i, \left\{ \bigcup_{i=1}^{k} V^i_x \middle| x \varepsilon \bigcup_{i=1}^{k} A_j \right\}, \bigcup_{i=1}^{k} C_i, \sum_{i=1}^{k} \mathrm{n_i} >$$

Union policy is the default. Alternatively, intersection policy can be used. When intersection policy is used, the global schema takes the tuple expression:

$$< \bigcap_{i=1}^{k} A_i, \left\{ \bigcup_{i=1}^{k} V^i_x \middle| x \varepsilon \bigcap_{j=1}^{k} A_j \right\}, \bigcup_{i=1}^{k} C_i, \sum_{i=1}^{k} \mathrm{n_i} >$$

07. CH creates conditional global schema for each site by dropping attributes and masking attribute values that were not originated from the site

B. Stage 2

08. CH communicates the conditional global schema to the k participating sites in the form: $< \{a_1 \ldots a_m\}, \left\{\left\{v_1^1, \ldots v_{d_1}^1\right\}, \ldots, \left\{v_1^m, \ldots v_{d_m}^m\right\}\right\}, \{c_0, \ldots c_n\} >$

09. Based on the attributes in the conditional global schema, each of the sites S_i generates uniform templates for the *crosstable matrices* for individual attributes. The crosstable matrix template for an attribute with m values $v_1, v_2, \ldots, v_m$ and n classes $c_1, c_2, \ldots, c_n$ takes the form:

	c_1 c_2 · · · c_n
v_1	
v_2	
⋮	
v_m	

The elements in the matrix templates are populated by the local site. Such a matrix for a given attribute is called the Unified Local Crosstable Matrix (ULCM) for that attribute. The crosstable matrices for all attributes at each site are sent to the CH.

10. The CH adds the k ULCM's (site-specific crosstable matrices) for each attribute u and creates Unified Global Crosstable Matrix (UGCM) for u. If CT_i represents the ULCM for an attribute u at site i and CT_g represents the UGCM for u, then $CT_g(x,y) = \sum_{i=1}^{k} CT_i(x,y)$. Let the UGCM for an attribute be as follows:

$$\begin{bmatrix} v_{11} & \cdots\cdots & v_{1n} \\ \vdots & \ddots & \vdots \\ \vdots & \ddots & \vdots \\ v_{m1} & \cdots\cdots & v_{mn} \end{bmatrix} \tag{1}$$

11. The weighted average impurity measure for the attributes is calculated and the attribute for split is chosen based on smallest value of weighted average impurity measure (highest gain) (Tan et al. 2006). For the matrix shown in (1), the weighted average impurity measure is given in (2).

$$\frac{-1}{\sum_{i=1}^{m}\sum_{j=1}^{n} v_{ij}} \left(\sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} v_{ij} \log_2 \frac{v_{ij}}{\sum_{k=1}^{n} v_{ik}} \right\} \right) \tag{2}$$

This formula is based on Quinlan's ID3 decision tree algorithm (Quinlan 1993). By default, only attributes that exist in all the G_i's are included in the impurity calculations. Alternatives are discussed in section 3.1 below.

12. To proceed to the next level of the decision tree, updated queries are generated for each branch of the decision tree, using the values of the attribute selected. Updated queries are sent to the same k sites selected in Step 02. The process repeats from step 03 with each of the queries until the classes are reached in the leaf nodes.

As mentioned earlier, G_i in each individual site i will remain the same throughout the decision tree creation process.

3.1 Complete and Relative Entropies

"It is an unfortunate fact of life that data often has missing attribute values" (Quinlan 1993). In ID3, the values of an attribute must be known for determining the outcome of a test at a decision node. This restriction was overcome in C4.5 (Quinlan 1993) by calculating the apparent gain based on instances with known values for a given attribute taking into consideration the fraction of such cases in the training set. We generalize this idea of missing and non-existing values to the case of hybrid data schemas as complete and relative entropies. We also take into consideration the strength of the known values for this generalization. If a substantial number of values are present for an attribute, the strength will be high. Such an attribute deserves a chance and qualifies to be considered for test.

When a query Q is applied on a data set, the collection of data instances that satisfies the query is referred to as "filtered data space" or simply "filtered space". If a given attribute u exists in all instances in the filtered space, we say that u spans the filtered space. Otherwise, u does not span the filtered space. When entropy is calculated using formula (2), an attribute may or may not span all of the filtered data space. Assume that site i (i=1,2,…,k) has t_i data instances and Q is a query to be applied. If $q(t_i)$ is the number of instances that satisfy Q in site i (i=1,2,…,k), the total number of instances globally satisfying Q is given by: $t_g = \sum_{i=1}^{k} q(t_i)$. For a given attribute, using the elements of UGCM in (1), if the value of t_g equals $\sum_{i=1}^{m}\sum_{j=1}^{n} v_{ij}$, the attribute spans the whole filtered space and the value computed using (2) is termed "complete entropy". If $\sum_{i=1}^{m}\sum_{j=1}^{n} v_{ij} < t_g$, then the attribute does not span the whole filtered space and the value computed using (2) is termed "relative entropy". The "relative entropy" is entertained for node splitting decision only if the ratio $r = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} v_{ij}}{t_g}$ is greater than a predefined threshold τ. i.e., $r>\tau$. Otherwise the attribute is ignored from node splitting decision. The ratio r is a measure of the strength of the attribute over the filtered space. The threshold τ is usually set as a percentage of the coverage of the attribute over the instances in the filtered space; typically a high percentage value (e.g. 90 %).

4 DIDT as a promotion algorithm

A promotion algorithm (Mathew and Obradovic 2013) incorporates finer resolution data analytics into coarser levels. Let there be k tiers of resolutions, starting with r_1 as the first tier resolution and r_k as the k^{th} tier resolution. A window at tier i is a set of r_i data points. The representative value (rv) for a window at tier i will be a value calculated using some designated function f_i on the data points in the corresponding window frame. The algorithm starts with 1st data point in a window of resolution r_1 and keeps on moving the data cursor, adding data points to a bin for the window, until the bin is full. When r_1 points accumulate in the bin corresponding to resolution r_1, a representative value (rv) is calculated using a function f_1 on the data points in

the current bin. Then this *rv* is promoted to the next bin corresponding to r_2. Once r_2 values accumulate in bin for r_2, the *rv* for this bin is calculated and promoted to the bin corresponding to r_3 and so on. After the *rv* for a bin is calculated and promoted to the next bin, the entries in the original bin are cleared. The formal version of the algorithm is given below:

Begin algorithm
variables: resolution,
$r_1, r_2, ..., r_k$ //k tiers of resolutions
$bin_{r1}, bin_{r2}, …., bin_{rk}$ //k bins corresponding to the resolutions
start:
{
initialize $bin_{r1}, bin_{r2}, …., bin_{rk}$ to be empty
while more data points are available in resolution r_1
add current data point to bin_{r1}
if (bin_{r1} is full)
i = 1, resolution = r_1
while $bin_{resolution}$ is full
rv = $f_i(bin_{resolution})$
add rv to $bin_{resolution->next}$
empty $bin_{resolution}$
i=i+1, $bin_{resolution}$=$bin_{resolution->next}$
end while
end if
end while
}
End algorithm

It is easy to see that DIDT is a promotion algorithm. Assume that there are *k* participating sites and there are *m* attributes among the data instances.

As can be seen from Fig. 3, r_1 has a resolution of *k* and each window corresponds to the local crosstable matrix of one of the *m* attributes. Thus there are *m* windows in this case. The *rv* for this level is the global crosstable matrix for the corresponding attribute. At the r_2 level, the

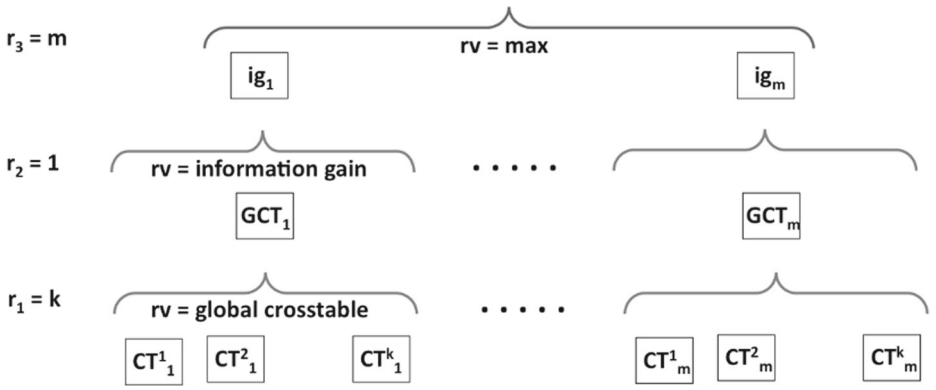


Fig. 3 Tiers in DIDT as a promotion algorithm

Table 1 Boolean Operations

Symbol	Operation	Precedence
!	negation	high
∧	conjunction	medium
∨	disjunction	low

information gain of each crosstable matrix is the *rv*. At the r_3 level, maximum value among the information gains is the *rv*.

5 Privacy model

We analyze 3 aspects of DIDT in this context; viz. the framework, privacy preservation and threat agents.

The framework includes the Clearing House (CH), the participating sites and their relationships. The CH is a trusted entity and acts as a central authority. The sites are vetted by the CH. This is very similar to federations (eg. caGrid) where the central authority has a credentialing process based on which membership is granted to institutions. Practical reasons for having the CH includes the vetting process and setting up a common code of ethics the member institutions should adhere to. This provides the administrative and technical controls to establish trust and data integrity.

Our interest is in the privacy of individuals as well as utility of the algorithm for practical use. The primary consideration in privacy preserving data mining is not disclosing Personally Identifiable Information (PII) similar to name, address, date of birth, etc. (Verykios et al. 2004). Since DIDT do not use raw data and does not need any PII attributes for prediction model building, this condition is satisfied. The secondary consideration is that of re-identification of the individual (Zheleva and Getoor 2007) or linking of microdata with publicly available information (Samarati 2001). Since DIDT does not require any raw data, no microdata is released and the problem related to linking of microdata for re-identification is eliminated. The third issue is related to targeted querying. This is the case of inference attacks (Silva et al. 2004) where focused queries are used to infer private information. In the case of DIDT, a query originating site cannot do a site-specific targeted query. The CH determines who participates in

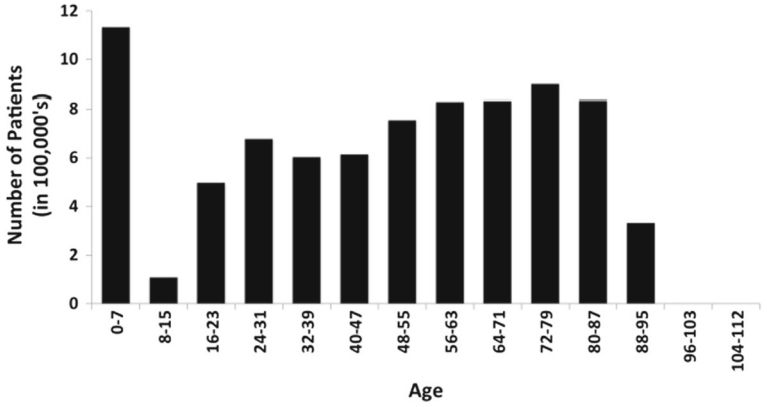


Fig. 4 Distribution of patient records based on age for NIS 2008 data

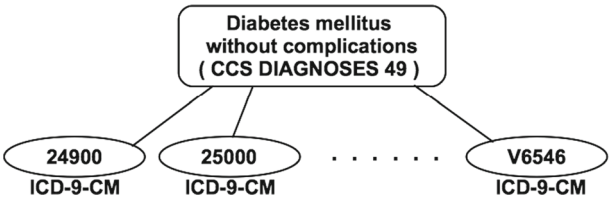


Fig. 5 Parent–child relationship between CCS code 49 and ICD-9-CM codes

the model building and the participating sites' credentials are not revealed to the query originator. Also, the sites selected by the CH are free to participate or not. This layer of isolation prevents targeted queries.

The main adversary in this model is an evesdropper. From a practical implementation point, asymmetric crypto keys can be used for one-to-one communication between the CH and the participating sites; thus thwarting evesdropping. For example, communications from the CH to site *x* can be encrypted using the public key of *x*.

6 Experiments

The experiments were oriented towards demonstrating the working of the DIDT algorithm in various business domains and verifying that the results are consistent with the centralized ID3 algorithm. For centralized runs, we used the well-known implementation of ID3 using weka opensource software (Hall et al. 2009). The implementation of DIDT was done in JAVA. The data instances in individual sites for DIDT were stored in neo4j (Huang and Dong 2013) graph databases in one-to-one mapping – one neo4j graph database per site. A graph database is well suited to represent heterogeneous records. The DIDT implementation in JAVA required one dedicated database per participating site. This was to ensure that the querying against individual databases and local crosstable generations are all working according to the published procedural steps of the algorithm. To aid in query processes, Lucene indexing (Bialecki et al. 2012) was made use of. In all experiments, a formal representation of queries using the Boolean operators in Table 1 was made use of.

One method of verifying a learning model is to use a set of training data to generate the model and use another set of data (called testing data) to verify the performance. This process is called cross-validation. For our distributed model, we use a leave-one-site-out cross-validation. In this type of cross-validations, if there are *n* sites, there will be *n* pairs of model building and validation. In one cross-validation, data from *n-1* sites will be used for training and data from the remaining site will be used for testing. For fair comparison between DIDT and ID3, we use the same (training, testing) sets for cross-validations in corresponding experiments.

Table 2 Results of the query: hospst = "CA" && (age > 12 && age < 20)

Algorithm	Number of sites	Cross-validation	Correctly classified	Accuracy
ID3	n/a	10	22629/23508	84.09 %
DIDT	83	10	22628/23508	84.09 %

Note that the attribute 'age' was not used in the classification as we were considering teenagers as belonging to one bin of age group

Table 3 Results for student data

Algorithm	Number of sites	Cross-validation	Correctly classified	Accuracy
ID3	n/a	6	1086/1194	90.95 %
DIDT	6	6	1091/1194	91.37 %

6.1 Experiments based on real data

6.1.1 Data from medical domain

The Nationwide Inpatient Sample (NIS) Database for 2008 was created by Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP). It contains discharge level information of all inpatients from a 20 % stratified sample of hospitals across USA. The 2008 NIS database has more than eight million "inpatient stay" records from 1056 hospitals. The distribution of patient records based on age is given in Fig. 4.

The distributions of male and female patients were 41.60 and 58.40 % respectively. Due to confidentiality laws, records with some very specific medical conditions and procedures (e.g. HIV/AIDS or abortion) are not released by certain hospitals.

There are up to 15 high level codes for diseases per data instance in the NIS 2008 data set. These are codes based on HCUP Clinical Classifications Software (CCS), developed by combining ICD-9-CM codes in a hierarchical fashion. For example, CCS code for diabetes mellitus without complications is 49. The Clinical Classifications Software for ICD-9-CM is a diagnosis and procedure categorization scheme where closely related ICD-9-CM codes are combined under a parent CCS code. There are a total of 259 CCS codes in all. The parent–child relationship with CCS Diagnoses 49 and its sibling ICD-9-CM codes is shown in Fig. 5.

Only 3 of the 12 children ICD-9-CM codes are shown in Fig. 5. The complete sibling ICD-9-CM codes are: 24900, 25000, 25001, 7902, 79021, 79022, 79029, 7915, 7916, V4585, V5391, and V6546. The largest percentages of records in the 2008 NIS data set were based on Essential Hypertension (CCS code 98) with 30.60 % and Coronary Atherosclerosis (CCS code 101) with 29.59 %.

The 259 CCS codes were represented as binary attributes. For a given hospitalization record, if a CCS code was present, the value of the corresponding binary attribute was set

Table 4 Distribution of data based on country

Country	# of instances
Brazil	1000
Chile	988
Guatemala	1305
India	1224
Kenya	1001
Nigeria	999
Philippines	1300
S. Africa	1137
S. Korea	1000
USA	1074

Table 5 Results for fully spanned attributes

Algorithm	Number of sites	Cross-validation	Correctly classified	Accuracy
ID3	n/a	10	6151/11028	55.78 %
DIDT	10	10	6228/11028	56.47 %

as 1 and if a CCS code was not present, the value of the corresponding binary attribute was set to 0. Thus, the 259 binary attributes represent the presence or absence of the corresponding CCS codes in a hospitalization record. In this experiment, we used 262 attributes for each hospitalization record. These were: Age, Race, Sex and the 259 binary attributes for CCS codes. The selection of these attributes was influenced by Khalilia et al.'s work (2011). Value for the attribute 'race' was missing from 5 states - Georgia, Illinois, Minnesota, Ohio and West Virginia – in the NIS 2008 data. Also, in some other hospitals, the attribute values for 'race' were missing from a portion of the records. In our experiment, we included only data instances for patient records that had all the attributes present.

The classification was done for a Californian teenage patient to be having "essential hypertension" (CCS Code 98) or not. The experiment was based on the query: hospst = "CA" && (age > 12 && age < 20). There were 83 hospitals that had patients with this criterion. Leave-one-hospital-out cross-validation would result in 83-fold cross-validations. So, in order to do 10-fold cross-validation, we used the method of banding hospitals into megahospitals (Mathew and Obradovic 2012). Three megahospitals were formed by randomly selecting 9 hospitals per band and 7 megahospitals were formed by randomly selecting 8 hospitals per band. Leave-one-megahospital-out cross-validations were performed and the results of DIDT as well as centralized ID3 decision trees are shown Table 2.

As can be seen from the results in Table 2, the DIDT algorithm gives result very consistent with the ID3 centralized result.

Table 6 Missing ratio of the additional 28 attributes

Attribute	Missing ratio	Attribute	Missing ratio
AIDS	2.70 %	ORG_REL7	86.28 %
ATTEND	3.78 %	POLITICS	2.70 %
CHANGE	77.64 %	PROSPER	2.70 %
CHRSTATE	1.21 %	Q5_A	17.56 %
CONVERT	7.00 %	Q5_B	17.56 %
GOD	93.17 %	Q5_C	17.56 %
HEALTH	2.70 %	Q5_D	17.56 %
ILLS	17.56 %	RAPTURE	17.56 %
MOSTIMP	7.00 %	REL_ALWS	1.09 %
ONLYWAY	17.56 %	RETURN	17.56 %
ORG_REL1	87.16 %	TONGUES	7.00 %
ORG_REL2	83.12 %	TONGUES2	19.84 %
ORG_REL5	89.45 %	TRUST8	7.00 %
ORG_REL6	88.30 %	WORK	7.00 %

Table 7 Results using supplemental attributes applying intersection policy

Entropy	Threshold	Cross-validation	Correctly classified	Accuracy
Complete	n/a	10	6380/11028	57.85 %
Relative	90 %	10	7612/11028	69.02 %
Relative	80 %	10	8170/11028	74.08 %

6.1.2 Data from higher-education domain

This experiment was based on student data from 6 schools. The UCI machine learning data repository has the data under the section 'Student Loan Relational Domain' (Frank and Asuncion 2010b). Each data instance had 7 attributes and classified as positive or negative. Positive means the student is not required to repay a student loan. For simplicity, the 'months of absence' attribute was coded into one of the categorical values 'low', 'med', and 'high' corresponding to the months of absence in the range 0–3, 4–5, and 6+ months (inclusive of the boundary values). Each attribute was treated as categorical in nature. The list of attributes and their categorical values used in the experiment are as follows:

sex {m,f}
absence {low, med, high}
units {1,2,3,4,5,6,7,8,9,10,11,12,13,14,15}
enlisted {none, armed, peace}
employed {yes, no}
bankruptcy {yes, no}
disabled {yes, no}

The data instances were distributed to 6 different sites corresponding to each school. Results of running DIDT and centralized ID3 on the full data sets are shown in Table 3.

As can be surmised from the results in Table 3, the DIDT results are very consistent with ID3 results. The difference in correct counts is due to the fact that multiple attributes can have same entropy and the selection of the attribute for node split is implementation-dependent.

6.1.3 Data from social domain

A multi-country survey commissioned by the Pew Forum on Religion and Public Life was done across 10 countries to investigate the religious, political and civic views of renewalists. The data (Spirit and Power 2006) consisted of 11,028 instances with 201 attributes. The distribution of the data among the countries is given in Table 4.

We treated the data instances as belonging to one of two classes: 'renewalists' or 'non-renewalists'. In the sections to follow, uppercase identifiers are attribute names in the data instances.

Table 8 Results of the query: SAMPLE=(1∨2) ∧ SATISFY=1 ∧ HARDWORK=1

Entropy	Threshold	Cross-validation	Correctly classified	Accuracy
Complete	n/a	10	1340/2333	57.44 %
Relative	90 %	10	1462/2333	62.67 %
Relative	80 %	10	1584/2333	67.90 %

a) Preprocessing:

The following preprocessing was done on the data set. CaseId$ (serial number), PSRAID (identification number), GPWGT (weight), and PENTWGT (weight) are attributes irrelevant for decision support. We distributed data based on COUNTRY, which provided naturally distributed data for 10 sites. And COUNTRY attribute cannot be used for tree building as otherwise leave-one-site-out cross validation cannot be performed. So, the 5 attributes CaseId$, PSRAID, GPWGT, PENTWGT, and COUNTRY were blocked from being released from each site using attribute constrains. Two attributes, AGE and CHANGE were continuous. So, they were categorized using concept hierarchy (Han and Fu 1994) into 'young','mid', and 'old'. All experiments were based on the query 'SAMPLE = 1 v 2', unless mentioned otherwise. The attribute SAMPLE was excluded from decision tree building.

b) Experiment based on fully spanned attributes:

In this experiment, only attributes that spanned over all the 11028 instances were included. Of the 201 attributes, 92 attributes met this criterion. Since the 5 attributes CaseId$, PSRAID, GPWGT, PENTWGT, and COUNTRY were constrained, that left 87 attributes. Ignoring the query attribute SAMPLE and the 'class' attribute, there were 85 attributes that contributed to the structure of the decision tree. The intersection policy and complete entropy were applied. Results are given in Table 5.

Ignoring the minor aberration due to the selection differences between the two algorithms for node splitting, here also we get results consistent between DIDT and centralized ID3.

6.2 DIDT-focused experiments

Since the consistency of DIDT with ID3 was demonstrated in the previous experiments, we focus on DIDT experiments for the rest of this section. Following experiments demonstrate additional capabilities of DIDT. The experiments were based on the pew survey data subjected to the same pre-processing outlined in the previous section. Some of the following experiments

Table 9 Cross-band for 'religious group' attribute

[a]	br	cl	gt	in	ke	ni	ph	za	kr	us
Q3BRA	1000	0	0	0	0	0	0	0	0	0
Q3CHI	0	988	0	0	0	0	0	0	0	0
Q3GUA	0	0	1305	0	0	0	0	0	0	0
Q3IND	0	0	0	1224	0	0	0	0	0	0
Q3KEN	0	0	0	0	1001	0	0	0	0	0
Q3NIG	0	0	0	0	0	999	0	0	0	0
Q3PHI	0	0	0	0	0	0	1300	0	0	0
Q3SAF	0	0	0	0	0	0	0	1137	0	0
Q3SKOR	0	0	0	0	0	0	0	0	1000	0
Q3US	0	0	0	0	0	0	0	0	0	1074

[a] *br* = Brazil, *cl* = Chile, *gt* = Guatemala, *in* = India, *ke* = Kenya, *ni* = Nicaragua, *ph* = Philippines, *za* = South Africa, *kr* = South Korea, *us* = USA

Table 10 Unification table for 'Religious group'

		Q3BRA	Q3CHI	Q3GUA	Q3IND	Q3KEN	Q3NIG	Q3PHI	Q3SAF	Q3KOR	Q3USA
1	Roman Catholic	1	1	1	3	1	1	11	1	1	2
2	Evangelical	2	2	2							
3	Afro-Brazilian	3									
4	Jehovah's Witness	4	3	3		9	6	4	7	5	9
5	Mormon		5	4				5		6	4
6	Buddhist	8								3	
7	No religion, …	10	11	8	10	8	9	10	10	8	8
8	Other religion	97	97			97	97		97		97
9	Kardecist	11									
10	Spiritism	12									
11	Don't know		98	98					98		98
12	Resused										99
13	Jewish		6	6					8		3
14	Mayan Traditional			5							
15	Hindu				1				4		
16	Muslim				2	4	3	6	3		6
17	Protestant				4	2	2	2	2	2	1
18	Syrian Orthodox				5						
19	Traditional tribal				11						
20	African Instituted					3					
21	African traditional					6			6		
22	Aladura Church						4		5		
23	Brotherhood of Cross & Star						7				
24	Iglesia ni Cristo							3			

Table 10 (continued)

		Q3BRA	Q3CHI	Q3GUA	Q3IND	Q3KEN	Q3NIG	Q3PHI	Q3SAF	Q3KOR	Q3USA
25	PIC							7			
26	Other Christian Group							8			
27	Other non-Christian group							9			
28	Confucianist								4		
29	Won-Buddhism								9		
30	Korea Soka Gakkai Intnl.									10	
31	Daesunjintihow									11	
32	Orthodox										5
33	Jain				6						

do not have ID3 equivalents because certain attributes in these experiments do not span all instances.

6.2.1 Experiment using supplemental attributes

The data attributes included in this experiment are those spanning all sites with at least one non-missing value in every site. This data set is different from the one in the previous experiment. Here, the attribute exists among all sites, but not necessarily in all instances within a site. i.e., certain attributes may have holes in some sites. An attribute may not span all 11028 samples in the first iteration. It can fully span a filtered space down the decision tree. There were 113 attributes in this intersection. Hence, 28 extra attributes are supplemented compared to the data set in previous section. These 28 additional attributes do not span all the 11028 instances. Table 6 shows the percentage of instances in which these attributes have missing values.

Results of applying intersection policy are shown in Table 7.

The slight improvement in accuracy for complete entropy in Table 7 compared to result in Table 5 is attributed to the additional features, which improves the precision of the decision tree. As seen from Table 7, relaxing the span of attributes over filtered space improved accuracy of the model. This is attributed to the fact that some features were able to produce better relative entropy in the node splitting selection.

6.2.2 Experiment based on the query: 'SAMPLE = (1∨2) ∧ SATISFY = 1 ∧ HARDWORK = 1'

Results of experiments using this query and applying intersection policy are as given in Table 8.

In this case also, improvement in accuracy is attributable to relative entropy. The shift in accuracy is following the same pattern as in Table 7.

6.3 Data unification

Due to the inherent hybrid nature of the data, it is possible to observe cross-bands of attributes. A cross-band exists if the matrix representing counts of non-missing values for local representations of an attribute across all the sites is a diagonal matrix. Each one of the local attributes that makes up a cross-band span one site only. In all other sites, the count will be 0. This is due to the fact that the values of a localized attribute are specific to the site. Table 9 shows the diagonal matrix of one such cross-band (for 'religious group').

Brazil had 1000 data instances and in all instances the Q3BRA attribute had non-missing values. Hence the (Q3BRA,Brazil) entry of the matrix is 1000. Q3BRA, Q3CHI, …, Q3USA are local representations for the attribute 'religious group'. In this case, we call 'religious group' the super attribute. In line with the local representations, we represent this super attribute as Q3 in the following discussions.

Table 11 Results of intersection policy applied on unified data

Entropy	Threshold	Cross-validation	Correctly classified	Accuracy
Complete	n/a	10	8792/11028	79.72 %
Relative	90 %	10	10049/11028	91.12 %
Relative	80 %	10	8904/11028	80.74 %

Table 12 Grouping values

		Q3BRA	Q3CHI	Q3GUA	Q3IND	Q3KEN	Q3NIG	Q3PHI	Q3SAF	Q3KOR	Q3USA
1	Roman Catholic	1	1	1	3	1	1	1	1	1	2
2	Evangelical	2	2	2							
3	Jehova's Witness	4	3	3		9	6	4	7	5	9
4	Mormon		5	4				5		6	4
5	Buddhist	8								3	
6	No Religion	10	11	8	10	8	9	10	10	8	8
7	Other Religion	97	97			97	97		97		97
8	Don't know/Refused		98	98					98		98,99
9	Jewish		6	6					8		3
10	Hindu				1				4		
11	Muslim				2	4	3	6	3		6
12	Protestant				4	2	2	2	2	2	1
13	African Traditional					6			6		
14	Other Christian Group	3			5	3	4,7	3,7,8	5		5
15	Other non-Christian Group	11,12		5	6,11			9		4,9,10,11	

Table 13 Results of intersection policy applied on unified data with value groupings

Entropy	Threshold	Cross-validation	Correctly classified	Accuracy
Complete	n/a	10	10253/11028	92.97 %
Relative	90 %	10	10287/11028	93.28 %
Relative	80 %	10	8733/11028	79.19 %

Unification is the process by which the local representations are combined to create a single super attribute using the superset of values of the local representations. Unification is in line with unified vocabulary among sites. It may or may not be possible to unify a cross-band, depending on whether the attributes are very specific to the country or not. In those cases where it is possible to unify a cross-band, the values of the local representations in all sites can be combined together by some unification process to create a super attribute that covers all sites with a range of the combined values. International organizations or agencies may have normalization tables that can be used for unification.

The data set had 8 cross-bands (7 spanned all 10 sites completely). Some of these cross-bands can be unified. 'Religious group' was one of the attributes with a cross-band effect that we unified. The following figure (Table 10) shows the table we used for unifying 'religious group' attribute among the 10 countries. In this table, Q3BRA, Q3CHI, …, Q3USA are country-specific (localized) representations for the 'religious group' attribute as used in the original data. Q3BRA uses the value 4 to represent 'Jehovah's Witness', whereas Q3GUA uses value 3, Q3KEN uses value 9, etc. So, we unify the value 'Jehovah's Witness' by assigning 4 as the common representative value among all sites. We continue this unification process for the other 32 values. Thus, unification is the process by which local representations are combined to create a super attribute by tagging the superset of values from the local representation.

We used the unified values from the unification table for Q3 and applied DIDT. Results are shown in Table 11.

The results in Table 11 are similar to the results from Tables 7 and 8 in that relaxing the span produced better relative entropies for node splitting decision. However, the threshold point for higher accuracy is observed to be shifting.

There are 33 values for the super attribute Q3 as seen from Table 10. Some of these values are sparse among sites. For example, the value "Brotherhood of Cross & Star" is confined to

Table 14 Cross-band effect of 'income'

INC_BRA	1000	0	0	0	0	00	0	0	0	0
INC_CHI	0	988	0	0	0	0	0	0	0	0
INC_GUA	0	0	1305	0	0	0	00	0	00	0
INC_IND	0	0	0	1224	0	00	0	0	0	0
INC_KEN	0	0	0	0	1001	0	0	0	0	0
INC_NIG	0	0	0	0	0	999	0	0	00	0
INC_PHI	0	0	0	0	0	0	1300	0	0	0
INC_SAF	0	0	0	0	0	0	0	1137	0	0
INC_SKOR	0	0	0	0	0	0	0	0	1000	0
INC_US	0	0	0	0	0	0	0	0	0	1074

Table 15 ppp factor used to unify income in units of US dollars

Country	ppp factor	Country	ppp factor
Brazil	1.562217325	Nigeria	82.40308843
Chile	364.677042	Philippines	22.3679723
Guatemala	3.085622277	S. Africa	4.854167618
India	13.64586631	S. Korea	775.051255
Kenya	35.85241959		

Nigeria and "Iglecia ni Cristo" is confined to Philippines. To improve the spread of values among sites to reduce sparsity, we grouped values that exist only in one site. The groupings were done primarily by combining singly isolated values (e.g. 'Jain') into one of two group values "Other Christian Group" and "Other non-Christian Group".

Table 12 shows the table after grouping values from the table in Table 10. As can be seen, the number of values reduced from 33 to 15 after this grouping.

The results based on these value groupings are shown in Table 13.

As seen from the results in Table 13 relaxing the span produced better accuracy, though not substantial. The threshold point for higher accuracy is following the same pattern as in Table 11.

6.4 Data unification for 'income' super attribute

INC_BRA, INC_CHI, …., INC_USA were the local representations of 'income'. The cross-band effect of these attributes is seen in Table 14.

We unified these using Penn World Table (Heston et al. 2009) of purchasing power parity (ppp) for all countries in the world. The ppp conversion values for 2006—the same year the data was collected—is shown in Table 15. The super attribute obtained by unifying the local representations was named INC. However, after the conversion, the income ranges were

Table 16 Mapping of local values to unified values for the super attribute 'income' for first set of 5 countries

INC_BRA	INC_CHI	INC_GUA	INC_IND	INC_KEN
1 → below	1 → below	1 → below	1 → below	1 → below
2 → poor	2 → poor	2 → poor	2 → poor	2 → average
3 → poor	3 → poor	3 → average	3 → average	3 → middle
4 → poor	4 → average	4 → average	4 → average	4 → rich
5 → poor	5 → average	5 → middle	5 → middle	98 → Don't know
6 → average	6 → middle	6 → middle	6 → middle	99 → Refused
7 → average	7 → middle	7 → upper-mid	7 → upper-mid	
8 → middle	8 → upper-mid	8 → upper-mid	8 → upper-mid	
9 → middle	9 → rich	9 → rich	9 → rich	
10 → upper-mid	10 → upper-mid	98 → Don't know	10 → rich	
11 → rich	98 → Don't know	99 → Refused	11 → upper-rich	
98 → Don't know	99 → Refused		98 → Don't know	
99 → Refused			99 → Refused	

Table 17 Mapping of local values to unified values for the super attribute 'income' for second set of 5 countries

INC_NIG	INC_PHI	INC_SAF	INC_SKOR	INC_USA
0 → below	1 → below	19 → below	1 → below	1 → below
1 → below	2 → poor	18 → poor	2 → poor	2 → poor
2 → poor	3 → average	17 → poor	3 → average	3 → average
4 → average	5 → upper-mid	15 → poor	5 → middle	5 → middle
5 → average	6 → rich	14 → average	6 → middle	6 → middle
6 → average	7 → upper-rich	13 → average	7 → middle	7 → upper-mid
7 → middle	99 → Refused	12 → average	8 → upper-mid	8 → rich
8 → middle		11 → average	9 → upper-mid	9 → upper-rich
9 → upper-mid		10 → middle	10 → upper-mid	98 → Don't know
10 → upper-mid		9 → middle	11 → rich	99 → Refused
11 → rich		8 → middle	12 → rich	
12 → rich		7 → middle	13 → upper-mid	
13 → upper-mid		6 → upper-mid	98 → Don't know	
98 → Don't know		5 → upper-mid	99 → Refused	
99 → Refused		4 → upper-mid		
		3 → rich		
		2 → rich		
		1 → Don't know		
		98 → Don't know		
		99 → Refused		

widely varied between countries. So, a binning process with bin selection was needed. Seven bins were created per individual country and individual income ranges within each country were allocated to specific bins (many-to-one mapping). Two values "Don't know" and "Refused" were left intact. The income-based bins were: "below poverty line", "poor", "average", "middle","upper-middle", "rich", "upper-rich".

Tables 16 and 17 shows the mappings from the original local data values to the super attribute values.

Applying the unified values for the super attribute 'income' to the data set, we get the results in Table 18.

These complete and relative entropy results are consistent with the results in Table 7. Hence, the unification of 'income' did not influence the accuracy of the classification.

The super attribute 'education' was obtained by the unification of local representations EDUC_BRA, EDUC_CHI, …, EDUC_USA. The unification was based on the

Table 18 DIDT results with thresholds, intersection policy, relative entropy and super attribute 'income'

Entropy	Threshold	Cross-validation	Correctly classified	Accuracy
Complete	n/a	10	6361/11028	57.68 %
Relative	90 %	10	7602/11028	68.93 %
Relative	80 %	10	8158/11028	73.98 %

Table 19 Query results of: *SAMPLE= (1∨2) ∧ (INC= rich ∨ EDUC= degree)*

Entropy	Threshold	Cross-validation	Correctly classified	Accuracy
Complete	n/a	10	1677/1858	90.26 %
Relative	90 %	10	1577/1858	84.88 %
Relative	80 %	10	1470/1858	79.12 %

widely accepted 8-4-4 (primary-secondary-college) format for years spent in the corresponding level. Unification of education did not produce better results. There were other attributes REG (region), ETH (ethnicity), and LAN (language) that had cross-band effects, but the values were specific to each country and so these attributes were not amenable to unification.

$$\textbf{\textit{Experiments based on the query: 'SAMPLE} = (1 \vee 2) \wedge \left(\textbf{\textit{INC} = \textit{rich} } \vee \textbf{\textit{EDUC} = \textit{degree}}\right)\textbf{'}$$

Unification of 3 attributes Q3 (religion), EDUC (education), and INC (income) were done at the same time on the distributed data and the query was run on the resulting data. 1858 instances satisfied the query. Results are shown in Table 19.

As evidenced from the results in Tables 7, 11 and 19, the maximum accuracy can be attained with complete or relative entropy. The level at which it is attained depends on the attributes distribution within the filtered space.

7 Conclusion

We have outlined an algorithm that can aid in decision-making when data is distributed among multiple sites and shipping raw data out from the individual sites is impractical due to regulatory and legal reasons. This algorithm for a distributed decision support system can protect personally identifiable information as well as assure data privacy. We have demonstrated that a distributed decision support intelligence tool in the form of a decision tree can be constructed using just the statistics related to the data distributed among the sites. We provided empirical validations for the key construct in our algorithm. Experiments done on various data sets from different business domains validate the results generated by the algorithm. Relative entropy was introduced to entertain attributes that span a minimum threshold globally to be considered for node split in the decision tree building. Unification of attribute values to create a set of common values among distributed sites was presented. It was empirically shown that the DIDT algorithm gives results that are consistent with the results generated by centralized ID3 algorithm. In summary, for DIDT algorithm, the data do not leave the sites and no rigid data schema structure is enforced on the collaborating sites. This makes it a viable option for building knowledge from sites that cannot disclose sensitive data records due to privacy concerns.

Acknowledgments This research was supported in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854.

References

Adam, N. R., & Wortman, J. C. (1989). Security control methods for statistical databases. *ACM Computing Surveys, 21*(4), 515–556.

Aggarwal, C. C., & Yu, P. S. (2008). *Privacy-preserving data mining: Models and algorithms*. New York: Springer Science+Business Media, LLC.

Allaert, F.-A., & Barber, B. (1998). Some Systems Implications of EU data protection directive. *European Journal of Information Systems, 7*(1), 1–4.

Bar-Or, A., Keren, D., Schuster, A., & Wolff, R. (2005). Hierarchical decision tree induction in distributed genomic databases. *IEEE Transactions on Knowledge and Data Engineering, 17*(8), 1138–1151.

Bialecki, A., Muir, R., & Ingersoll, G. (2012). Apache Lucene 4. *ACM SIGIR Workshop on Open Source Information Retrieval* (pp. 17–24). Portland, OR, USA.

Brand, R. (2002). Microdata protection through noise addition. *Inference Control in Statistical Databases*. Lecture Notes in Computer Science, Vol. 2316. Springer-Verlag, Berlin-Heidelberg.

Canetti, R. (1998). Security and composition of multi-party cryptographic protocols. *Journal of Cryptography, 2000*(13), 143–202.

Caragea, D., Silvescu, A., & Honavar, V. (2004). A framework for learning from distributed data using sufficient statistics and its application to learning decision trees. *International Journal on Hybrid Intelligent Systems, 1*(1–2), 80–89.

Chow, C., & Mokbel, M. F. (2011). Trajectory privacy in location-based services and data publication. *ACM SIGKDD Explorations: Special Issue on Privacy in Mobility Data Mining, 13*(1), 19–29.

Cieslak, D. A., Hoens, T. R., Chawla, N. V., & Kegelmeyer, W. P. (2012). Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery, 24*(1), 136–158.

Courtright, C. G. (2001). Criteria for developing clinical decision support systems. *14th IEEE Symposium on Computer-Based Medical Systems* (pp. 270 – 275). Bethesda, MD, USA.

Dinur, I., & Nissim, K. (2003). Revealing information while preserving privacy. *22nd ACM Symposium on Principles of Database Systems (PODS)* (pp. 202–210). San Diego, CA, USA.

Du, W., & Atallah, M.J. (2001). Secure multi-party computation problems and their applications: A review and open problems. *New Security Paradigms Workshop* (pp. 11–20). Cloudcroft, NM, USA.

Dwork, C. (2006). Differential privacy. *33rd International Colloquium on Automata, Languages and Programming (ICALP)* (pp. 1–12). Venice, Italy.

Frank, A., & Asuncion, A. (2010a). SPECT heart data set, UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml/datasets/SPECT+Heart

Frank, A., & Asuncion, A. (2010b). Student loan relational data set, UCI machine learning repository. Irvine, CA: University of California, School of Information and Computer Science. http://archive.ics.uci.edu/ml/datasets/Student+Loan+Relational

Friedman, A., & Schuster, A. (2010). Data mining with differential privacy. *16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 493–502). Washington D.C., USA.

Fu, Y. (2001). Distributed data mining: An overview. *Newsletter of the IEEE Technical Committee on Distributed Processing*. Spring 2001, 5–9.

Giannella, C., Liu, K., Olsen, T., & Kargupta, H. (2004). Communication efficient construction of decision trees over heterogeneously distributed data. *Fourth IEEE International Conference on Data Mining* (pp. 67–74). Brighton, UK.

Goldreich, O. (1998). Secure multi-party computation. Available at http://www.wisdom.weizmann.ac.il/~oded/pp.html.

Goldstein, D. E. (2000). *e-Healthcare: Harness the power of internet e-commerce & e-care* (pp. 417–418). Gaithersberg: Aspen Publishers Inc.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutermann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations, 11*(1), 10–18.

Han, J., & Fu, Y. (1994). Dynamic generation and refinement of concept hierarchies for knowledge discovery in databases. *AAAI'94 Workshop Knowledge Discovery in Databases (KDD'94)* (pp. 157–168). Seattle WA, USA.

Heston, A., Summers, R., & Aten, B. (2009). *Penn World Table Version 6.3. Center for International Comparisons of Production, Income and Prices*. USA: University of Pennsylvania.

Huang, H., & Dong, Z. (2013). Research on architecture and query performance based on graph database Neo4j. *3rd International Conference on Consumer Electronics, Communications and Networks (CECNet)* (pp. 533–536). Xianning, China.

Inan, A., Saygyn, Y., Savas, E., Hintoglu, A. A., & Levi, A. (2006). Privacy preserving clustering on horizontally partitioned data. *22nd International Conference on Data Engineering Workshops*, 95. Atlanta, GA, USA.

Kantarcioglu, M. (2008). A survey of privacy-preserving methods across horizontally partitioned data. *Advances in Database Systems, 34*, 313–335.

Kantarcioglu, M., & Clifton, C. (2004). Privacy-preserving distributed mining of association rules on horizontally partitioned data. *IEEE Transactions on Knowledge and Data Engineering, 16*(9), 1026–1037.

Kantarcioglu, M., Nix, R., & Vaidya, J. (2009). An efficient approximate protocol for privacy-preserving association rule mining. *13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)* (pp. 515–524). Bangkok, Thailand.

Karthikeyan, G., & Pais, P. (2010). Clinical judgment and evidence-based medicine: time for reconciliation. *Indian Journal of Medical Research, 132*(5), 623–626.

Khalilia, M., Chakraborty, S., & Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Medical Informatics and Decision Making*, http://www.biomedcentral.com/1472-6947/11/51

Khoshgoftaar, T. M. (2005). Identifying noise in attributes of interest. *Fourth International Conference on Machine Learning Applications* (pp. 55–60). Boca Raton, FL, USA.

Kumbhar, M. N., & Kharat, R. (2012). Privacy preserving mining of association rules on horizontally and vertically partitioned data: A review paper. *12th International Conference on Hybrid Intelligent Systems (HIS)*, (pp. 231–235). Pune, India.

(2000). Privacy-preserving data mining. *Advances in Cryptology – CRYPTO '00, Lecture Notes in Computer Science*. Springer-Verlag, Berlin-Heidelberg. 1880, 36–53.

Lindell, Y., & Pinkas, B. (2009). Secure multiparty computation for privacy-preserving data mining. *The Journal of Privacy and Confidentiality, 1*(1), 59–98.

Mathew, G., & Obradovic, Z. (2010). Vocabularies in collaboration channels. *6th International Conference on Collaborative Computing: Networking, Applications and Work Sharing* (pp. 1–5). Chicago, IL, USA.

Mathew, G., & Obradovic, Z. (2011a). Constraint graphs as security filters for privacy assurance in medical transactions. *2nd ACM Conference on Bioinformatics, Computational Biology and Biomedicine* (pp. 502–504). Chicago, IL, USA.

Mathew, G., & Obradovic, Z. (2011b). A privacy-preserving framework for distributed clinical decision support. *1st IEEE International Conference on Computational Advances in Bio and medical Sciences* (pp. 129–134). Orlando, FL, USA.

Mathew, G., & Obradovic, Z. (2012). Distributed privacy preserving decision system for predicting hospitalization risks in hospitals with insufficient data. *Machine Learning in Health Informatics Workshop: International Conference on Machine Learning Applications - ICMLA* (pp. 178–183). Boca Raton, FL, USA.

Mathew, G. & Obradovic, Z. (2013). Improving computational efficiency for personalized medical applications in mobile cloud computing environment. *IEEE International Conference on Healthcare Informatics, The First Workshop on Mobile Cloud Computing in Healthcare* (pp. 535–540). Philadelphia, PA, USA.

Moret, B. M. E. (1982). Decision trees and diagrams. *ACM Computing Surveys, 14*(4), 593–623.

Navathe, S., Ceri, S., Wiederhold, G., & Dou, J. (1984). Vertical partitioning algorithms for database design. *ACM Transactions on Database Systems, 9*(4), 680–710.

Park, B-H., & Kargupta, H. (2003). Distributed data mining: Algorithms, systems and applications. In N. Ye (Ed.), *The handbook of data mining* (pp. 341–358). Lawrence Erlbaum Associates.

Pinkas, B. (2002). Cryptographic techniques for privacy-preserving data mining. *SIGKDD Explorations, 4*(2), 12–19.

Quinlan, J. R. (1986). Introduction to decision trees. *Machine Learning, 1*, 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo: Morgan Kaufmann Publishers.

Rockwell, R. C., & Abeles, R. P. (1998). Sharing and archiving data is fundamental to scientific progress. *Journal of Gerontology Series B: Psychological Sciences and Social Sciences., 53*(1), S5–S8.

Samarati, P. (2001). Protecting respondents' identities in Microdata release. *IEEE Transactions on Knowledge and Data Engineering, 13*(6), 1010–1027.

Silva, J. C. D., Klusch, M., Lodi, S., & Moro, G. (2004). Inference attacks in peer-to-peer homogeneous distributed data mining. *16th European Conference on Artificial Intelligence (ECAI)* (pp. 450–454). Valencia, Spain.

Spirit and Power: A 10-Country Survey of Pentecostals. (2006). Available at: http://www.thearda.com/Archive/Files/Descriptions/PENTEC.asp

Sweeney, L. (2010). Data Sharing Under HIPAA: 12 Years Later. Advance HIT Project. White paper 1006. USA: Harvard University.

Tan, P., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson Addison Wesley.

Vaidya, J., & Clifton, C. (2003a). Privacy-preserving k-means Clustering over Vertically Partitioned Data. *ACM SIGKDD International Conference on Knowledge Discovery and Data* (pp. 206–215). Washington, DC, USA.

Vaidya, J., & Clifton, C. (2003b). Leveraging the "Multi" in secure multi-party computation, *ACM Workshop on Privacy in the Electronic Society* (pp. 53–59). Washington, DC, USA.

Vaidya, J., & Clifton, C. (2005). Privacy-preserving decision trees over vertically partitioned data. *Lecture Notes in Computer Science*, Springer, Berlin-Heidelberg. 3654, 139–152.

Vaidya, J., & Clifton, C. (2009). Privacy-preserving Kth element score over vertically partitioned data. *IEEE Transactions on Knowledge and Data Engineering, 21*(2), 253–258.

Verykios, V. S., Bertino, E., Fovino, I. N., Provenza, L. P., Saygin, Y., & Theodoridis, Y. (2004). State-of-the-art in privacy preserving data mining. *SIGMOD Record, 33*(1), 50–57.

Vest, J. R., & Gamm, L. D. (2010). Health information exchange: persistent challenges and new strategies. *Journal of American Medical Association, 17*(3), 288–294.

Wu, Y., Jiang, X., & Ohno-Machado, L. (2012). Grid Binary LOgic REgression (GLORE): building shared models without sharing data. *Journal of American Medical Informatics Association, 19*(5), 758–764.

Xu, Z. (2011). Classification of privacy-preserving distributed data mining protocols. *6th International Conference on Digital Information Management* (pp. 337–342). Melbourne, Australia.

Yao, A. C. (1986). How to generate and exchange secrets. *27th IEEE Symposium on Foundations of Computer Science* (pp. 162–167). Toronto, Canada.

Yu, H., Vaidya, J., & Jiang, X. (2006). Privacy-preserving svm classification on vertically partitioned data. *Advances in Knowledge Discovery and Data Mining, 3918*, 647–656.

Zheleva, E., & Getoor, L. (2007). Preserving the privacy of sensitive relationships in graph data. *Privacy, Security and Trust in KDD, First ACM SIGKDD International Workshop (PinKDD),* (pp. 153–171). San Jose, CA, USA.