Netw Model Anal Health Inform Bioinforma
DOI 10.1007/s13721-013-0041-y

ORIGINAL ARTICLE

Dynamic distributed predictive learning models that preserve privacy for hospitals with insufficient labeled data

George Mathew · Zoran Obradovic

Received: 17 June 2013 / Revised: 9 August 2013 / Accepted: 10 August 2013

Abstract A prediction model built dynamically using patient data from multiple hospitals can serve as a tool for suggestive knowledge in clinical decision support. Such a tool that accommodates queries based on attributes of interest is helpful in building a targeted model from multiple hospitals when a local clinical data repository does not have sufficient number of records to draw conclusions from. However, because of privacy concerns and legal ramifications, hospitals are reluctant to divulge raw medical records. Hence, mechanisms to build distributed prediction models using just the statistics of patient data are attractive. Distributed ID3-based decision tree (DIDT) algorithm is such a prediction model builder. In this study, we analyze National Inpatient Sample data for 3 years and demonstrate that DIDT can be used to help collaboratively build better predictive models when hospitals have insufficient number of records for good local models. Using 261 attributes for model building, we showed that collaborating hospitals with less than 100 cases of hospitalizations for a targeted disease were able to achieve good improvement in accuracies for predicting hospitalization collectively using a distributed model compared to local models. When relying on local models for predicting risks for sample diseases, more patients were misclassified and some local patients could not be classified. Our collaborative model effectively reduced misclassification providing accurate early diagnostics to additional patients. The profile of hospitals with sufficiently large number of patient records was explored to identify local models with specific characteristics that can serve the needs of hospitals with insufficient data.

Keywords Distributed decision making · Privacy preserving prediction model · Hospitalization risk prediction

1 Introduction

Practices in medical domain are “characterized by much judgmental knowledge” (van Melle 1978), and consequently suggestive models that can help in decision making are valuable to clinical practitioners. Survey results have also confirmed that physicians are interested in such decision support systems (Sittig et al. 2006). First generation clinical decision support systems (Buchanan and Shortliffe 1984; Bobrow et al. 1986) were rule-based and static in nature. They could not learn from the body of new patient data generated over periods of time. Building knowledge from opportunistic data is the hallmark of data mining techniques. Data mining algorithms have shown to be helpful with building models in domain-specific applications. Identifying patients at risk for targeted communications (Khalilia et al. 2011) have been accomplished by applying data mining methods. Other prediction models of recent interest are related to emergency admissions (Li et al. 2012) and hospital readmission costs (Kansagara et al. 2011). Personalized medicine has also benefited from data mining techniques (Wegener et al. 2013). In our study, we target privacy-preserving classification models built dynamically from distributed databases for predicting hospitalization risk.

G. Mathew (✉) · Z. Obradovic
Center for Data Analytics and Biomedical Informatics, Temple University, 324 Wachman Hall, 1805 N. Broad Street, Philadelphia, PA 19122, USA
e-mail: George.Mathew@temple.edu

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

Published online: 28 August 2013

Because of independent existence of hospitals under different business administrations, the collection of patient data is geographically distributed among various hospitals. As a consequence of privacy concerns and regulatory implications, collecting raw patient data from distributed hospitals to a central location is not practical (Loukides et al. 2010). Patterns of similar logistic issues in other business domains have led to the emergence of privacy preserving distributed data mining (PPDDM) (Xu 2011) as a recent area of research interest. From a clinical practice perspective, there is interest in building decision support systems that can harness the power of collective intelligence from multiple hospitals using the power of Internet (Sittig et al. 2008). Data privacy can be accomplished in distributed environments by employing cryptographic protocols. Privacy preserving distributed clustering has been demonstrated using Healthcare data (Elmisery 2010) in this manner. A simpler privacy preserving distributed model building mechanism can be based on algorithms that use just the statistics of the patient data from multiple hospitals. Such algorithms do not require sophisticated cryptographic infrastructures. Prediction models built in a distributed fashion are valuable tools in medical practice. In certain clinical situations, the local patient database may not have sufficient number of records of a certain diagnosis to garner intelligence from. In these cases, dynamically mining the collective distributed space of similar hospitals in a collaborative fashion can possibly lead to a quite useful decision making model. For example, a particular patient may be an outlier in the physician's practice and so it would help to obtain information relevant to diagnosis and treatment from external hospitals. Another scenario is the case of a patient with rare disease. Since the information to be mined is seeded by the attributes of the patients at hand, a mechanism to query based on the "attributes of interest" (Khoshgoftaar and Van Hulse 2005) will be helpful. The objective of our study is to help draw conclusions on a certain diagnosis using shared statistics from multiple hospitals when there are not enough samples locally. A hypothesis explored in this study is that mining the collective distributed data space of similar hospitals in a collaborative fashion can possibly lead to developing a better decision making model when the collaborating hospitals do not have sufficient number of instances to make good decisions on their own. Based on this premise, we explored hospitals in the Nationwide Inpatient Sample (NIS 2013) data sets for the years 2007–2009, each of which had less than 100 patient records having a targeted disease. For those hospitals, we built the local models and compared them to the distributed model built using distributed ID3-based decision tree (DIDT) (Mathew and Obradovic 2011) algorithm. The distributed model using just the statistics of data provided noticeable improvement in accuracy over average local model accuracies.

DIDT is a simple algorithm that produces a decision tree identical to the one produced on an equivalent centralized data aggregation. A decision tree (Moret 1982) is a data structure that represents the paths of traversals in a decision-making process for classification problems. ID3 (Quinlan 1986) is a centralized decision tree building algorithm and is used as the reference algorithm in DIDT. One of the techniques from C4.5, where possible values are allocated among different groups with one outcome for each group, is used in this study. Other tests from C4.5 can be incorporated, if need be. We deal only with categorical attributes in this study and so ID3 base is sufficient. Since DIDT uses only statistics of data from the distributed hospital databases, it is a valuable tool in privacy preserving distributed decision-making. DIDT has a built-in mechanism to search the distributed databases using logical constructs based on specified attributes of interest. This search facility helps identify precisely the targeted data instances from the distributed pool of databases. For example, if a patient with a specific set of symptoms and vital signs is an outlier in the local database, these attributes of interest can be used to seed the initial distributed search.

The equivalency of DIDT to centralized tree building is theoretically provable. This means that the model built by DIDT algorithm by learning from distributed data sets is provably exact (Caragea et al. 2004) with respect to its centralized counterpart. Thus, there is no loss of fidelity in the results produced by our distributed algorithm DIDT. This is attractive compared to privacy preserving algorithms similar to differential privacy (Dwork 2006) that introduces noise to the statistics and hence introduce distortion to the results.

It is a common practice for small hospitals to associate with larger hospitals for better bargaining power in business world and for leveraging access to additional medical resources. In such instances, it is possible that the bigger hospital may have sufficient number of patient records to build a prediction model that the smaller hospital can use. We explored this idea to understand the characteristics of such data sets that can help predictions in smaller hospitals.

2 NIS data (2007–2009)

The Nationwide Inpatient Sample (NIS) databases for years 2007–2009 was created by Agency for Healthcare Research and Quality (AHRQ) Healthcare Cost and Utilization Project (HCUP). Published NIS databases contain discharge level information of all inpatients from a 20 % stratified sample of hospitals across USA. Each data

Table 1 Details of patient records in the NIS 2007–2009 data sets

Years	Total number of patient records	Total number of hospitals	Male (%)	Female (%)
2007	8,043,415	1,044	41.26	58.74
2008	8,158,381	1,056	41.60	58.40
2009	7,810,762	1,050	41.92	58.08

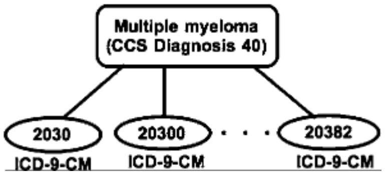

Fig. 1 Parent-child relationship between CCS code 40 and ICD-9-CM codes

instance in these data sets represents an "inpatient stay record". Because of local/state confidentiality laws, some specific medical conditions or procedures (e.g., HIV/AIDS) are not released by certain hospitals. Individual records in the NIS data are de-identified. That is, they do not carry personally identifiable information (e.g., name or home address). Hence, they provide a vertical partition of attributes that are ideal candidates for use in a privacy preserving distributed decision support model. The variations in instances between hospitals give a real world setting to study distributed algorithms. The number of records and hospitals as well as the distribution of male and female patients among the NIS 2007–2009 data sets are given in Table 1.

High-level disease codes in the NIS data are based on HCUP Clinical Classifications Software (CCS), developed by combining ICD-9-CM codes in a hierarchical fashion. For example, CCS code for Multiple myeloma is 40. The CCS 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 259 CCS codes in all. There are up to 15 CCS codes for diseases per data instance in the NIS 2007–2008 data sets, while the NIS 2009 data set has up to 25 CCS codes per instance. The parent–child relationship with CCS diagnosis 40 (Multiple myeloma) and its sibling ICD-9-CM codes is shown in Fig. 1.

Only 3 of the 8 ICD-9-CM codes that make up CCS 40 are shown in Fig. 1. The complete list of ICD-9-CM sibling codes is: 2030, 20300, 20301, 20302, 2038, 20380, 20381, and 20382.

The distribution of patient records among the 2007–2009 NIS data sets based on age is given in Fig. 2.

The distribution of patient records based on race is given in Fig. 3.

The distribution of race in Fig. 3 is based on the uniform HCUP race code. The values corresponding to these codes are:

1 -	white
2 -	black
3 -	hispanic
4 -	asian or pacific islander
5 -	native american
6 -	others

The distribution of the five most common specific comorbidities among the patient records over the years 2007–2009 were as given in Table 2. Comorbidities have been studied for valuable clues using prediction models from data mining techniques (Himes et al. 2009) and clustering models from statistical methods (Yang et al. 2013).

Our study was focused on patients with "Diabetes mellitus without complications" (CCS code 49) in NIS 2009 data sets, on patients with "Chronic obstructive pulmonary disease and bronchiectasis" (CCS code 127) in NIS 2008 data set and on "Congestive heart failure; nonhypertensive" (CCS Code 108) in NIS 2007 data set.

3 Related works

The NIS data sets have been used in various medical studies with a statistical approach. Age-related cholecystectomy (Kuy et al. 2011) analysis was done using NIS data from 1996–2001. Factors affecting length of hospital stay in connection with mouth cellulitis (Kim et al. 2012) were analyzed using NIS 2008 data. Hospitalization costs and post discharge follow-up care costs associated with meningococcal disease were studied (Davis et al. 2011) making use of 2005 NIS data. These studies were using traditional statistical instruments with a centralized data model. Studies using data mining techniques on public data sets were also published. Support vector machine prediction was used for diabetes-related hospitalization (Yu et al. 2010). A recent study provided an enhancement to the support vector machine-recursive feature elimination (SVM-RFE) mechanism to optimally estimate disease risk based on 2008 and 2009 NIS data (Stiglic et al. 2012). Random forest technique for predicting disease risks was applied by Khalilia et al. (2011) on the NIS 2005 data. An improved prediction model over this work, using fuzzy membership based on ICD-9 codes later appeared in the literature (Popescu and Khalilia 2011). All these data mining techniques address classification problem and are

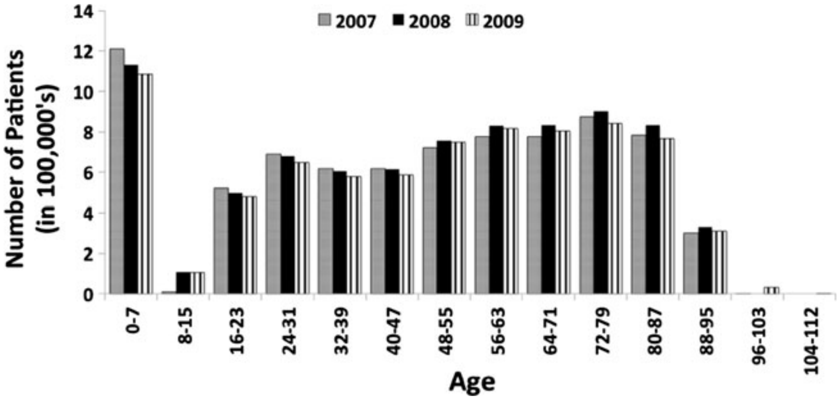

Fig. 2 Distribution of patient records for NIS 2007–2009 data sets based on age

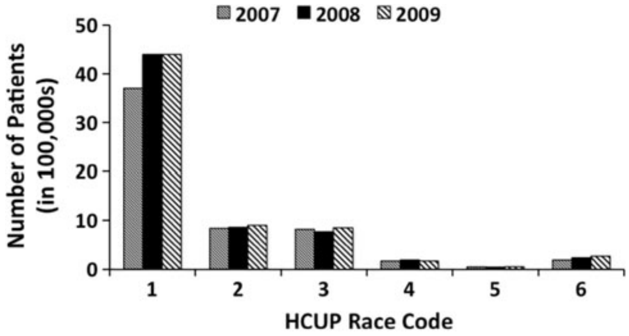

Fig. 3 Distribution of patient records for NIS 2007–2009 data sets based on uniform HCUP race code

based on centralized data architecture. Centralized mechanisms require all data instances to be available at a central site. Decision trees form a group of popular classification algorithms because of their simplicity. The first serial decision tree building algorithm was proposed by Quinlan. Parallel decision tree building algorithms (Jin and Agrawal 2003) to speed up model building also appear in literature. In real life, the patient records are distributed among clinical databases in various hospitals. Because of this natural distribution of patient data among hospitals, a distributed data mining technique (Park and Kargupta 2003) would align well with the distributed data topology. Distributed data mining can use existing distributed computing infrastructure similar to grids (Luo et al. 2007). Privacy preserving data mining is relevant in the context of our work because of the privacy issues related to patient data. Privacy preserving support vector machine (svm) (Yu et al. 2006) can be trained in a distributed fashion. But this work was based on vertically partitioned data. A recent work in distributed privacy preserving model building is on Logistic regression (Wu et al. 2012). Though these distributed algorithms preserve patient privacy, they require identical data schema among the participating sites and they do not have mechanisms to dynamically specify the attributes of interest. Distributed hierarchical decision tree (DHDT) (Bar-Or et al. 2005) is a distributed decision tree building algorithm that focuses on high dimensional data for reducing communication costs and takes advantage of the correlations among attributes. Distributed ID3-based decision tree (DIDT) is a distributed decision tree building algorithm that builds a classification model and does not assume any correlations between attributes. DHDT assumes identical data schema among participating hospitals, while DIDT can accommodate non-identical data schema. In addition, DIDT has a built-in search mechanism to initiate a query based on the attributes of interest.

We present theory and experimental results of two methods of predictive model building that can be used by

Table 2 Prevalence rates of comorbidities among NIS 2007–2009 data sets

CCS code	Description	2007 (%)	2008 (%)	2009 (%)
98	Essential hypertension	29.11	30.60	31.20
101	Coronary atherosclerosis	27.90	29.59	31.18
55	Fluid and electrolyte disorders	21.05	21.80	
53	Disorders of lipid metabolism	17.40	19.43	
259	Residual codes	15.57	18.57	
106	Cardiac dysrhythmias			16.80
108	Congestive heart failure			15.14
49	Diabetes mellitus without complications			14.88

hospitals when they do not have sufficient number of samples locally to build good prediction models for diagnosis. First method uses DIDT to build prediction models collaboratively with other hospitals (Mathew and Obradovic 2012) and preliminary empirical exploration of this method was done on a single data set - NIS 2009. In this study, we extend our empirical investigation to two additional NIS patient data sets from years 2007 and 2008. The second method introduced here uses the prediction model from a hospital with enough samples and having certain signature. We characterize the profile of such hospitals with large number of data instances whose prediction models can help hospitals handicapped with insufficient data. We show empirically that these models can provide good accuracy compared to the models in the first method. In addition, we assess the statistical significance of the models based on the two methods and show their acceptability.

4 Methodology

DIDT is a privacy preserving distributed decision tree building algorithm. It uses the count of values of attributes across classes among patient data distributed among hospitals to build the prediction model. This information is captured by the data structure known as crosstable matrix (Caragea et al. 2004). If an attribute a takes values $v_1, v_2, \ldots, v_m$ spread across classes $c_1, c_2, \ldots, c_n$ among the instances in a given patient database, the (x, y)th element of the crosstable matrix corresponding to a is the count of data instances having class label c_y for which attribute a has value v_x. The template for crosstable matrix corresponding to attribute a having the characteristics mentioned above takes the form (Mathew and Obradovic 2012):

$$\begin{array}{c|ccc} & c_1 & \cdots & c_n \\ \hline v_1 & \square & & \\ \vdots & \square & & \\ v_m & \square & & \end{array} \tag{1}$$

The crosstable matrix formats across all participating hospitals corresponding to each attribute are maintained uniformly. The sum of the crosstable matrices from individual hospitals is called global crosstable matrix. For a given attribute, the global crosstable matrix represents the complete distribution of the values among all classes. The global crosstable matrices can be used to calculate information gains. The attribute that gives maximum gain is picked to generate the next down-level branches of the decision tree. Assume that the global crosstable matrix for attribute a based on template (1) is as follows (Mathew and Obradovic 2011):

$$\begin{bmatrix} b_{11} \cdots\cdots b_{1n} \\ \vdots \quad \ddots \quad \vdots \\ \vdots \qquad \ddots \; \vdots \\ b_{m1} \cdots\cdots b_{mn} \end{bmatrix} \tag{2}$$

Then, the formula for computing the weighted average impurity measure for attribute a is (Mathew and Obradovic 2011):

$$\frac{-1}{\sum_{i=1}^{m}\sum_{j=1}^{n} b_{ij}} \left(\sum_{i=1}^{m} \left\{ \sum_{j=1}^{n} b_{ij} \log_2 \frac{b_{ij}}{\sum_{k=1}^{n} b_{ik}} \right\} \right) \tag{3}$$

The global crosstable matrix for each attribute is used to calculate its weighted average impurity measure. The attribute with the highest gain (smallest weighted average impurity measure) is chosen for test (Tan et al. 2006). Once an attribute is chosen for test, the logical expression representing the path of traversal from root to each of the new nodes is constructed using Boolean operations. These logical expressions are used for searching the data sets to globally identify attributes to be considered for the new set of crosstable matrices and eventual down level node splits. These steps are repeated until leaf nodes are reached.

The DIDT algorithm uses a centralized agent called Clearing House to mediate between the query originator and the distributed hospitals. A step-by-step working of the DIDT algorithm is outlined below:

```
1. A query Q by a medical practitioner is sent to the Clearing House (CH).
2. CH sends the query Q to k hospitals $S_1, \ldots, S_k$.
3. for (i = 1 to k) {

      $D_i$ be the local patient database in hospital $S_i$;
      $I_i$ = set of instances matching query Q in $D_i$;
      $A_i$ = set of all attributes among instances in $I_i$;
      $C_i$ = set of all classes in $I_i$;
      for each x e $A_i$ {
      $V_x^i$ = set of values for x;
      }
      The metadata tuple in the form
   $\langle A_i, \{V_x^i \mid x \in A_i\}, C_i, |I_i| \rangle$        (4)
   is sent to CH;
      }
4. CH aggregates the k tuple expressions in (4) to create a global schema.
5. for (i = 1 to k) {
    $S_i$ receives global schema in the format
                  n n           o     n              o o
$\backslash \ \{a_1;\ldots a_m\};\quad v^1_1;::v^1_{d_1}\quad;::;\quad v^m_1;::v^m_{d_m}\quad;\{c_0;::c_t\}[$

from CH;
    for each a e{$a_1$, …, $a_m$} {
    Template $T_a$ is created in the format (1);
    crosstable matrix is computed using layout $T_a$;
    crosstable matrix is sent to CH;
    }
    }
```

6. For each attribute a e{a_1, …, a_m}, CH sums up the site-specific crosstable matrices to create global crosstable matrix for a.
7. The weighted average impurity measure for the attributes are calculated using (3) and the attribute with smallest value of weighted average impurity measure (highest gain) is chosen for test.
8. To proceed to the next level of the decision tree, updated queries are generated for each branch of the decision tree, based on values of the attribute selected for test.
9. The process repeats from step 2 with each of the new queries until the classes are reached in the leaf nodes.
10. The CH sends final decision tree to the query originator.

Cross-validation is done by leave-one-hospital-out method. In this method, data from one hospital are used for testing, while data from all other hospitals are used for training. When a large number of hospitals participate, the leave-one-hospital-out cross-validation method will lead to a large number of cross-validations. Hence, we modified the cross-validation method in the original DIDT algorithm to accommodate a varied form such that the number of cross-validations can be kept at 10. In this modified format, a set of hospitals are combined together to create a logical mega-hospital. Then a mega-hospital can be left out for testing. Mega-hospital building was implemented by randomly selecting appropriate number of hospitals without replacement so that these mega-hospitals provide a partition of all participating hospitals. For example, when there are 500 participating hospitals, a leave-one-hospital-out cross-validation will necessitate 500-fold cross-validations. On the other hand, 10 mega-hospitals formed by picking 50 hospitals at a time randomly without replacement can be used for 10-fold cross-validations using the mega-hospitals. In our study, this modified version of DIDT was used for distributed model building. Weka (Hall et al. 2009) open source software was used for building local models with 10-fold cross-validations. Age attribute was categorized using a binning process (Elomaa et al. 1999). A range of 8 years (starting with ages 0–7) was used for one bin.

The 2009 NIS data set had up to 25 CCS codes per hospitalization record, while the 2007–2008 NIS data sets had up to 15 CCS codes per record. All the 259 CCS codes were represented as binary attributes in each instance for experiments. For a given hospitalization record, the value of the binary attribute corresponding to a CCS code was set as 1 or 0 depending on the presence or absence of the CCS code in the record. In our classification, 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 (2011) work. Values for the attribute 'race' were missing from some states. In our study, we excluded hospitals from these states. Details related to this information are given in Table 3.

Even in the non-excluded hospitals from other states, the attribute values for 'race' were missing from a portion of the records. In these cases, we included only data instances for patient records that had all the attributes present.

5 Experiments

5.1 Pre-processing

The SPSS load program from the AHRQ-HCUP web site was used to load the NIS 2007–2009 data files into SPSS Statistics software (Ver. 19) from IBM. From SPSS, we exported data records as comma separated values (csv) based text files. These csv files were parsed using PERL scripts and corresponding arff format files were created. 'arff' is a data input format used by Weka software.

The experiments were done in a simulated distributed environment. The way we implemented DIDT in JAVA, the code requires one dedicated (or self-contained) database per hospital for patient data. This ensures that the querying for matches against individual databases and local cross table generations for attributes are all working in accordance with the published procedural steps of the DIDT algorithm. Since we were using the NIS data sets and not live patient databases from real hospitals, the patient records corresponding to each hospital within the NIS data

Table 3 Information regarding missing attribute 'race' in NIS 2007–2009 patient records

Years	Number of instances with all attributes present	States with race attribute missing
2007	5,807,267	GA, IL, KY, ME, MN, NV, OH, OR, WA, WV
2008	6,520,461	GA, IL, MN, OH, WV
2009	6,614,593	MN, NC, OH, WV

set were extracted and loaded into individual databases in one–one mapping—one Neo4j (2013) graph database (Cook and Holder 2007; Aggarwal and Wang 2007) per hospital. The experiments were done as a simulation using this group of databases that provided the virtual distributed environment of hospitals. Using graph model for the data framework helps capture the underlying structure of clinical data in a very natural way. The symptoms associated with a patient visit were represented as the labeled vertices of a graph. A graph database is well suited to represent heterogeneous records. Lucene (2013) indexing was used for text indexing within the neo4j databases. Decision tree building on mega-hospitals and individual hospitals were done using Weka software.

5.2 Baseline experiment using NIS 2009 dataset

In this section, we present the published baseline experiment (Mathew and Obradovic 2012). We studied the problem of classifying patients with or without “Diabetes mellitus without complications” using NIS 2009 data set. Only hospitals with all 262 attributes present were taken into account. There were 902 such hospitals. Local models were built for these 902 hospitals using 10-fold cross-validations. The distribution of the resulted accuracy ranges is shown in Table 4.

As observed from Table 4, 23 hospitals had local models with less than 60 % accuracy. To further evaluate the distribution in this range so as to identify possible improvements, patient records from these 23 hospitals were tallied into ranges as shown in Table 5.

As can be observed from Table 5, the prevalence of hospitals in this group had less than 100 patient records. So, we decided to focus on this group of 11 hospitals, each of which had less than 100 patient records. Of these 11 hospitals, 5 could not build local models and 2 had less than 3 records. Using less than 3 records from one hospital can possibly lead to reverse-identifying individual patient(s) in a distributed system. Hence, we decided to leave out these hospitals from our study. Thus, we targeted the 9 hospitals having less than 100 patient records. We generated local decision trees with 10-fold cross-validations. The results are as shown in Table 6.

Table 4 Spread of prediction accuracies across 902 hospitals having diabetes records (Mathew and Obradovic 2012)

Accuracy ranges	Count of hospitals
Could not build classifier	5
Below 50 %	1
50–60 %	17
60–70 %	86
70–80 %	411
80–90 %	353
90–100 %	29

Table 5 Distribution of hospitals having\ 60 % accuracies in local prediction models

Count of patient records	Count of hospitals
1–100	11
101–200	2
201–300	4
301–400	4
401–500	1
501–600	1

Table 6 Spread of prediction accuracies across hospitals having\ 100 patient records with diabetes feature (Mathew and Obradovic 2012)

Count of patient records	Count of hospitals	Local prediction accuracy
1–10	3	–
25–50	2	56–60 %
51–75	3	48.48–58.06 %
76–100	1	57.5 %

Average local prediction accuracy among the 9 target hospitals was calculated using:

$$\frac{\text{count of correctly classified instances in all 9 hospitals}}{\text{total instances in all 9 hospitals}}$$
$$= 53{:}08\,\%$$

DIDT algorithm performed on the same set of hospitals yielded an accuracy of 63 %, an improvement of 9.92 %. For comparison with equivalent centralized model, data from all 9 hospitals were combined centrally and decision tree was built on this data using the same cross-validation splits as the one used by DIDT to avoid cross-validation mismatch. This resulted in an accuracy of 64.07 % and is recorded in Table 7 (second row).

It is seen from Table 7 that DIDT gives empirical result close to its centralized equivalent. The aberration in the result is due to the fact that multiple attributes can have identical information gains and so any one of them can be chosen for a given node split. Consequently, the distributed trees are not necessarily identical to one another. However, the big advantage of DIDT over its centralized equivalent is that no raw patient record is required from the hospitals—only statistics of the patient data is needed. The centralized tree building requires raw patient data from all hospitals in a central location and is costly in terms of data communication costs as well as in terms of data privacy.

Table 7 Prediction accuracies for 9 hospitals having\ 100 records

Method	Accuracy (%)	Count of incorrectly diagnosed patients
Average of local predictions	53.08	159[a]
DIDT	63	138
Centralized equivalent	64.07	136

[a] 16 patients from the 3 hospitals that could not build local models were excluded from this count

Table 8 Improvements in accuracies contributed by DIDT among hospitals having\ 1,000 records (Mathew and Obradovic 2012)

Count of patient records	Count of hospitals	Increase in accuracy contributed by DIDT (%)
1–100	9	9.92
1–250	12	4.92
1–500	19	1.49
1–1,000	20	0.37

The statistics of not harming patients by improved diagnosis using DIDT is shown in the last column of Table 7. These results show another advantage of using DIDT. The number of patients incorrectly classified is quite less with DIDT in comparison to what local models do on their own, even after excluding the 16 patients that could not be classified by the local models from the count and accounting for them in DIDT. The collaborative distributed model we used reduced the misclassification from 159 to 138 effectively providing accurate early diagnostics to 37 additional patients.

As observed from the values in Table 8, employing DIDT resulted in an improvement of 9.92 % in accuracy. We postulated that this method gives the best net improvement in accuracy. To validate this hypothesis, improvements in accuracies were computed when hospitals collaborate with hospitals having higher number of similar patient records. This was done using DIDT algorithm to generate the corresponding decision trees on related patient records from the 9 hospitals augmented by the hospitals in the corresponding tier. The resulting improvements in accuracies for various tiers are shown in Table 8.

It is seen from Table 8 that the net improvement in accuracy was best when the disadvantaged hospitals with less than 100 patient records used DIDT to build a distributed prediction model. Collaboration of hospitals with insufficient number of data with hospitals having larger number of records did not contribute to improve the accuracy substantially. A visual representation of this trend is shown in Fig. 4.

Hospitals having large number of patient records to build local models tend to fare better on their own since they can build prediction models specific to their patients.

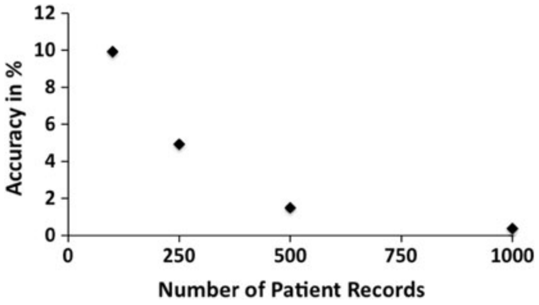

Fig. 4 Plot of accuracies across increasing resolutions among hospitals

5.3 Experiments based on NIS 2007–2008 data sets

In line with the baseline experiment, we conducted two other classification studies. First problem was classifying patients with or without “congestive heart failure” (CCS code 108) using NIS 2007 data set. Second problem was classifying patients with or without “chronic obstructive pulmonary disease and bronchiectasis” (CCS code 127) using NIS 2008 data set. The experiments were based on data from hospitals with all 262 attributes present. There were 757 hospitals in the NIS 2007 data set and 855 hospitals in the NIS 2008 data set with data instances having non-missing values for age, sex and race. In line with the baseline experiment, local models for these hospitals with less than 100 patient records were generated with 10-fold cross-validations and those hospitals with less than 60 % accuracy were identified. The results are shown in the fourth column of Table 9. Note that in following tables, result from baseline experiment is also added for easy comparison.

Applying DIDT with leave-one-hospital-out cross-validation, we got the results shown in the last column of Table 9.

It is observed from last two columns of Table 9 that the improvements in accuracies are consistent with the baseline experiment. The statistics of not harming patients by improved diagnosis is shown in Table 10.

As can be seen from Table 10, the improvements in statistics of not harming patients are also consistent with the baseline results.

5.4 Dimension reduction

The next experiment was oriented towards reducing the dimension of the patient data. Feature selection (Van Hulse et al. 2012) and feature reduction (Mathew and Obradovic 2013) are common techniques for pre-processing high dimensional data. It was observed that some comorbidities do not exist among the aggregated data set. Hence, attributes corresponding to these symptoms were eliminated,

Table 9 Accuracies for hospitals with\ 100 records using local models and DIDT

NIS data year	CCS code	Count of hospitals	Average local model accuracy (%)	DIDT accuracy (%)
2007	108	6	51.59	61.90
2008	127	5	59.09	66.09
2009	49	9	53.08	63

Table 10 Distribution of incorrect diagnosis of patients for hospitals with\ 100 records

NIS year	Using local models	Using DIDT
2007	54[a]	47
2008	72[a]	57
2009	159[a]	138

[a] Excluding patients from hospitals without local models

Table 11 Distribution of prediction accuracies after dimension reduction for hospitals with\ 100 diagnosis-related records

NIS data year	Reduction in number of features	Prior DIDT (%)	Lower dimension DIDT (%)
2007	129	61.90	61.90
2008	126	66.09	65.51
2009	91	63	63

resulting in much reduced dimensions as shown in Table 11.

Comparing the columns in Table 11 for prior DIDT and lower dimension DIDT, it is observed that the accuracy with reduced dimension remains very close to the prior even after considerable reduction in dimension.

5.5 Experiments using prediction models of other hospitals

A natural follow-up question is, whether there are local models from hospitals with large number of instances that can serve as good prediction models for hospitals with insufficient data. If such models exist, what are their profiles? Note that a predication model does not contain any patient information and so it can be passed to another site without any privacy violations. To explore this idea, we considered hospitals with high accuracy ([95 %), high number of positive instances and high area under curve (AUC). The interest in higher number of positives is because patients have multiple comorbidities and hence positive class instances for the diagnosis of interest tend to be much lower.

First, we identified hospitals with greater than 95 % accuracy. These hospitals were ranked based on two factors: the number of positive instances and the AUC. Weights were assigned based on ranking and the weighted average was calculated between number of positive instances and AUC. We only considered the top 5 ranks. For rank r, the weight was 1– (r - 1) 9 0.10 = 1.10 – r 9 0.10. For example, in the NIS 2007 data set for CCS Code 108, the distribution of accuracy, number of positive instances and the AUC of top 4 hospitals are as shown below in Table 12.

Table 12 Distribution of number of positives and AUC for hospitals with [95 % accuracy in NIS 2007 data set

Hospid	Accuracy (%)	Number of positives	AUC	Weighted average
12323	96.43	549	0.67	0.9
36194	97.63	385	0.66	0.8
6558	95.12	312	0.76	0.9
6577	95.48	238	0.69	0.8

For the entry with hospid 12323, the rank for number of positives is 1, while the rank for AUC is 4. Hence, the weighted average is (1.10 – 1 9 0.10 ? 1.10 – 3 9 0.10)/2, which is 0.9. The model with the highest weighted average is chosen. For the purpose of this discussion, we call this model the weighted model. Using this method, we identified hospitals in the previous 3 experiments with accuracy [95 %. From these lists, hospitals were ranked and the weighted models were selected. Using these weighted prediction models to classify instances in the hospitals with insufficient data for years 2007–2009 resulted in accuracies shown in Table 13.

Based on the results in Table 13, we observe that the weighted model gives accuracy better than the collaboratively built DIDT model.

5.6 Statistical significance

In this section, we statistically evaluate the significance of the models developed in Sect. 5.5 compared to the ones in Sect. 5.3.

Assume there are k cross-validations. Let d_i be the difference in error rates between the decision trees in Sects. 5.3 and 5.5 at ith cross-validation. Since we do a leave-one-hospital-out cross-validation, the numbers of instances are not consistent among the training/testing sets across cross-validations. To compensate for this aberration, we use weighted mean:

Table 13 Prediction accuracies of DIDT and weighted models for hospitals with\ 100 diabetes-related records

NIS data year	CCS code	DIDT accuracy (%)	Weighted model accuracy (%)
2007	108	61.90	63.27
2008	127	66.09	66.66
2009	49	63	66.49

Table 14 Confidence intervals

NIS data	CCS code	1 *	r weighted	d^{t}_{cv}
2007	108	0.22	0.13	(- 0.04,0.48)
2008	127	0.10	0.19	(- 0.31,0.51)
2009	49	0.10	0.11	(- 0.10,0.30)

$$1^{*} = \frac{P_{i=1}^{k} d_i w_i}{P_{i=1}^{k} w_i}$$

Here,

$$w_i = \frac{\text{number of records for testing in ith iteration}}{\text{total number of records in the whole system}}$$

The weighted variance is calculated using the formula:

$$r^{2}_{\text{weighted}} = \frac{P_{1=1}^{k} w_i (d_i - 1*)^2}{P_{i=1}^{k} w_i}$$

Then the confidence interval d^{t}_{cv} is determined by:

$d^{t}_{cv} = 1 * \square\ t_{(1\text{-}a),k\text{-}1} \cdot r_{\text{weighted}}$

Here $t_{(1\text{-}a),k\text{-}1}$ is the t-distribution co-efficient with k-1 degrees of freedom and confidence level (1-a). Using these formulas, the confidence intervals for the models in Table 13 at 95 % confidence level are shown in Table 14.

As can be seen from Table 14, in all cases, the confidence interval span zero and so the error rates are not statistically significant.

6 Conclusion

Using NIS data for 2007–2009, we demonstrated that the DIDT algorithm can be employed to the advantage of hospitals that do not have enough information to build a local decision support model to collaboratively build a distributed model using just the statistics of data from such hospitals. The DIDT algorithm does not require patient data from participating hospitals. It improves the overall accuracy of a classification model and provides the disadvantaged hospitals with a classification model that otherwise would not be at their disposal. The error in diagnosis is reduced by DIDT. Though DIDT is a general-purpose distributed decision making algorithm, we demonstrated this algorithm could be used to address a very specific problem. We studied the model building in the case of predicting hospitalization based on three diseases. Since this methodology has no dependency on the disease per say it can be applied to build a classification model for any disease. We also improved efficiency of the leave-one-hospital-out cross-validation method in DIDT implementation to include the megahospital concept by banding together hospitals. The local models of hospitals with high accuracy, high AUC and high number of positive instances provided slightly better results compared to the collaboratively built DIDT models. The dimension reduction process produced nearly identical results compared to the original data.

Acknowledgments We acknowledge partial financial support from grant #FA9550-12-1-0406 from the U.S. Air Force Office of Scientific Research (AFOSR) and the Defense Advanced Research Project Agency (DARPA). This research was also supported in part by the National Science Foundation through major research instrumentation grant number CNS-09-58854. Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality provided data used in this study.

References

Aggarwal CC, Wang H (2007) Mining and managing graph data. Wiley-Interscience, Hoboken

Bar-Or A, Keren D, Schuster A, Wolff R (2005) Hierarchical decision tree induction in distributed genomic databases. IEEE Trans Knowl Data Eng 17(8):1138–1151

Bobrow DG, Mittal S, Stefik MJ (1986) Expert systems: perils and promise. Commun ACM 29(9):880–894

Buchanan BG, Shortliffe EW (1984) Rule based expert systems: the MYCIN experiments in the Stanford heuristic programming project. Addison-Wesley, Reading, Massachusetts

Caragea D, Silvescu A, Honavar V (2004) A framework for learning from distributed data using sufficient statistics and its applications to learning decision trees. Int J Hybrid Intell Syst 1(1–2):80–89

Cook DJ, Holder LB (2007) Mining graph data. Wiley Interscience, Hoboken

Davis KL, Misurski DA, Miller JM, Bell TJ, Bapat B (2011) Cost of acute hospitalization and post-discharge follow-up care for meningococcal disease in the United States. Hum Vaccin 7(1):96–101

Dwork C (2006), Differential privacy. In: proceedings of 33rd International colloquium on automata, languages and programming, pp 1–12

Elmisery AM (2010) Privacy preserving distributed learning clustering of healthcare data using cryptography protocols. In: proceedings of 34th annual IEEE computer software and applications conference workshops, pp 140–145

Elomaa T, Rousu J (1999) General and efficient multisplitting of numerical attributes. Mach Learn 36(3):201–244

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

Himes BE, Dai Y, Kohane IS, Weiss ST, Ramoni MF (2009) Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records. J Am Med Inform Assoc JAMIA 16(3):371–379. doi:10.1197/jamia.M2846

Jin R, Agrawal G (2003) Communication and memory efficient parallel decision tree construction. In: proceedings of 3rd SIAM international conference on data mining (SDM), pp 119–129

Kansagara D, Englander H, Salanitro A, Kagen D, Theobald C, Freeman M, Kripalani S (2011) Risk prediction models for hospital readmission. JAMA 306(15):1688–1698

Khalilia M, Chakraborty S, Popescu M (2011) Predicting disease risks from highly imbalanced data using random forest. BMC Med Inform Decis Mak 11:51. doi:10.1186/1472-6947-11-51

Khoshgoftaar TM, Van Hulse J (2005) Identifying noise in an attribute of interest. In: proceedings of 4th international conference on machine learning and applications, pp 55–62

Kim MK, Nalliah RP, Lee MK, Allareddy V (2012) Factors associated with length of stay and hospital charges for patients hospitalized with mouth cellulitis. Oral Surg Oral Med Oral Pathol Oral Radiol 113(1):21–28

Kuy S, Sosa JA, Roman SA, Desai R, Rosenthal RA (2011) Age matters: a study of clinical and economic outcomes following cholecystectomy in elderly Americans. Am J Surg 201(6): 789–796

Li J, Guo L, Handly N, Mai AA, Thompson DA (2012) Semantic-enhanced models to support timely admission prediction at emergency departments. Netw Model Anal Health Bioinform 1(4):161–172. doi:10.1007/s13721-012-0014-6

Loukides G, Denny JC, Malin B (2010) The disclosure of diagnosis codes can breach research participants' privacy. J Am Med Inform Assoc 17(3):322–327

Lucene (2013) Lucene project from Apache foundation. http://lucene.apache.org

Luo P, Lu K, Shi Z, He Q (2007) Distributed data mining in grid computing environments. Future Gener Comp Sys 23(1):84–91

Mathew G, Obradovic Z (2011) A privacy-preserving framework for distributed clinical decision support. In: proceedings of the 1st IEEE international conference on computational advances in bio and medical sciences, pp 129–134

Mathew G, Obradovic Z (2012) Distributed privacy preserving decision system for predicting hospitalization risks in hospitals with insufficient data. In: proceedings of ICMLA. pp 178-183

Mathew G, Obradovic Z (2013) Auto-reduction of features for containing communications costs in a distributed privacy-preserving clinical decision system. In proceedings of 3rd IEEE international conference on computational advances in bio and medical sciences

Moret BME (1982) Decision trees and diagrams. ACM Comput Surv 14(4):593–623

Neo4j (2013) Home page for neo4j graph database. http://neo4j.org Accessed June 2013

NIS (2013) Overview of the Nationwide Inpatient Sample data. http://www.hcup-us.ahrq.gov/nisoverview.jsp Accessed June 2013

Park B, Kargupta H (2003) Distributed data mining: algorithms, systems and applications. In: Ye N (ed) The handbook of data mining. Lawrence Erlbaum Associates, New Jersey, pp 341–358

Popescu M, Khalilia M (2011) Improving disease prediction using ICD-9 ontological features. In: 2011 IEEE international conference on fuzzy systems, pp 1805–1809

Quinlan JR (1986) Induction of decision trees. Mach Learn 1:81–106

Sittig DF, Krall MA, Dykstra RH, Russell A, Chin HL (2006) A survey of factors affecting clinician acceptance of clinical decision support. BMC Med Inform Decis Mak 6:6. doi:10.1186/1472-6947-6-6

Sittig DF, Wright A, Osheroff JA, Middleton B, Teich JM, Ash JS, Cambell E, Bates DW (2008) Grand challenges in clinical decision support. J Biomed Inform 41:387–392. doi:10.1016/j.jbi.2007.09.003

Stiglic G, Pernek I, Kokol P, Obradovic Z (2012) Disease prediction based on prior knowledge. In: proceedings of ACM SIGKDD workshop on health informatics, in conjunction with 18th SIGKDD conference on knowledge discovery and data mining

Tan P, Steinbach M, Kumar V (2006) Introduction to data mining. Pearson Addison Wesley, Boston, p 160

Van Hulse J, Khoshgoftaar TM, Napolitano A, Randall Wald (2012) Threshold-based feature selection techniques for high-dimensional bioinformatics data. Netw Model Anal Health Bioinform 1(1–2):47–61. doi:10.1007/s13721-012-0006-6

van Melle W (1978) MYCIN: a knowledge-based consultation program for infectious disease diagnosis. Int J Man Mach Stud 10(3):313–322

Wegener D, Rossi S, Buffa F, Delorenzi M, Ruping S (2013) Towards an environment for data mining based analysis processes in bioinformatics and personalized medicine. Netw Model Anal Health Bioinform 2(1):29–44. doi:10.1007/s13721-013-0022-1

Wu Y, Jiang X, Kim J, Ohno-Machado L (2012) Grid binary logistic regression (GLORE): building shared models without sharing data. J Am Med Inform Assoc 19(5):758–764

Xu Z (2011) Classification of privacy-preserving distributed data mining protocols. In: proceedings of sixth international conference on digital information management, pp 337–342

Yang M, Yang F, Oyang Y (2013) Application of density estimation algorithms in analyzing co-morbidities of migraine. Netw Model Anal Health Bioinform 2(2):95–101. doi:10.1007/s13721-013-0028-8

Yu H, Vaidya J, Jiang X (2006) Privacy-preserving svm classification on vertically partitioned data. Adv Knowl Discov Data Min 3918:647–656

Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ (2010) Application of support vector machine modeling for prediction of common diseases: the case of diabetes pre-diabetes. BMC Med Inform Decis Mak. doi:10.1186/1472-6947-10-16