Distributed Privacy-Preserving Decision Support System for Highly Imbalanced Clinical Data

GEORGE MATHEW and ZORAN OBRADOVIC, Temple University

When a medical practitioner encounters a patient with rare symptoms that translates to rare occurrences in the local database, it is quite valuable to draw conclusions collectively from such occurrences in other hospitals. However, for such rare conditions, there will be a huge imbalance in classes among the relevant base population. Due to regulations and privacy concerns, collecting data from other hospitals will be problematic. Consequently, distributed decision support systems that can use just the statistics of data from multiple hospitals are valuable. We present a system that can collectively build a distributed classification model dynamically without the need of patient data from each site in the case of imbalanced data. The system uses a voting ensemble of experts for the decision model. The imbalance condition and number of experts can be determined by the system. Since only statistics of the data and no raw data are required by the system, patient privacy issues are addressed. We demonstrate the outlined principles using the Nationwide Inpatient Sample (NIS) database. Results of experiments conducted on 7,810,762 patients from 1050 hospitals show improvement of 13.68% to 24.46% in balanced prediction accuracy using our model over the baseline model, illustrating the effectiveness of the proposed methodology.

Categories and Subject Descriptors: J.3 [**Life and Medical Systems**]: — *Medical information systems*; K.6.5 [**Management of Computing and Information Systems**]: Security and Protection

General Terms: Algorithms

Additional Key Words and Phrases: Distributed decision support, privacy preserving frameworks, clinical decision support system, privacy

ACM Reference Format:
Mathew, G. and Obradovic, Z. 2013. Distributed privacy-preserving decision support system for highly imbalanced clinical data. *ACM Trans. Manage. Inf. Syst.* 4, 3, Article 12 (October 2013), 15 pages.
DOI: http://dx.doi.org/10.1145/2517310

1. INTRODUCTION

Evidence-based clinical practice is influenced much by judgmental knowledge [Sim et al. 2001]. Consequently, a decision support system that can provide a medical practitioner with suggestive knowledge is a valuable tool of the trade. Survey results [Sittig et al. 2006] have affirmed the interest of physicians in such systems. The first wave of clinical decision support systems [Bobrow et al. 1986; Buchanan and Shortliffe 1984] were self-contained and rule-based. They had preconfigured rules to help make decisions on specific modalities. Because of the static nature of these rules, these systems could not harness the power of opportunistic data resident in multiple hospitals. With the availability of distributed systems [Oster et al. 2007; Warren et al. 2007]

This work is supported in part by the National Science Foundation under major research instrumentation grant number CNS-09-58854435060.
Authors' address: G. Mathew and Z. Obradovic, Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA; email: {George.Mathew, Zoran.Obradovic}@temple.edu.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2013 ACM 2158-656X/2013/10-ART12 $15.00
DOI: http://dx.doi.org/10.1145/2517310

and protocols [Dolin et al. 2001], it became possible to harness the power of clinical databases from multiple sources. Distributed classification algorithms [Caragea et al. 2004; Mathew and Obradovic 2011] make it possible to do decision making dynamically from multiple hospitals. The data could be retrospective or real time [Kansagara et al. 2011]. These systems use just the statistics of data from individual hospitals and hence preserve patient privacy. Clinical decision support systems can take advantage of data mining techniques to dynamically build the required information model. One class of widely used algorithms for such activity is classification. The goal of classification is to separate patient records into distinct classes. Classification has been shown to be effective in analyzing effects of various factors on diseases including clinical and demographic variables [Risch 2000]. The effectiveness of classification is dependent on the distribution of class attributes in the data. Classification algorithms typically assume a uniform distribution that provides balanced class data. Real-world classification problems are known to have imbalanced data. The attributes used for clinical model building are usually linked directly to the vital signs or physical characteristics of the patient and the diagnosis codes associated with the medical condition. There could also be comorbidities associated with the patient at the time of visit. In those situations where the conditions of the patient are rare occurrences, there could be very few or none of similar records in the local database to draw conclusions from. However, it is very likely that such records exist in other hospitals, though few in number in individual hospitals. We present a system where statistics of the data from various sites can be used to detect the class imbalance problem and build a classification model that gives much improved balanced accuracy compared to the baseline model. Our model building uses oversampling of the rare class data and as-needed undersampling of the abundant class instances. The model we build is an ensemble of decision trees. A decision tree [Moret 1982] provides a representation for the paths of traversals in the decision making process associated with a classification problem. The first serial decision tree-building algorithm was proposed by Quinlan [1986]. Algorithms for parallel construction [Jin and Agrawal 2003] as well as distributed construction of decision trees appear in the literature. Our decision model is a voting machine that is an ensemble of multiple decision trees built in a distributed manner. In the next section we provide the literature review. In Section 3, we outline our methodology and the system design. Computational complexities and communication costs are summarized in Section 4. The details of the Nationwide Inpatient Sample (NIS) data set are described in Section 5 and we present our experimental results based on the NIS data in Section 6.

2. RELATED WORKS

Data mining techniques for classification have been previously applied to public medical data. Support Vector Machine [Yu et al. 2010] was used for diabetes related hospitalization prediction. An enhancement to the Support Vector Machine—the Recursive Feature Elimination (SVM-RFE) method, has been proposed [Stiglic et al. 2012] to optimally estimate disease risk. These algorithms work well in balanced data sets. The class imbalance problem in classification is well documented [Japkowicz 2000]. A class imbalance occurs when the number of data instances that belongs to one class is very large while very few data instances belong to the other class. For example, a local clinical database may have one or zero prior record of syncope (ICD-9M code 7802). Algorithms that address class imbalance problems [Ertekin et al. 2007; Khalilia et al. 2011] appear in literature. Khalilia et al. apply Random Forest technique for disease prediction. An improved model using fuzzy membership based on ICD-9 codes was later proposed [Popescu and Khalilia 2011]. However, all of these require raw

patient data and cannot be used in privacy preserving systems. Also, these algorithms are centralized in nature and not designed for distributed systems. DHDT (Distributed Hierarchical Decision Tree) [Bar-or et al. 2005] is a distributed classification algorithm that preserves data privacy. However, the focus of the algorithm is on high-dimensional data and for reducing communication costs. It achieves efficiency by taking advantage of the correlations among attributes. DIDT (Distributed Id3-based Decision Tree) [Mathew and Obradovic 2011] is a privacy-preserving distributed algorithm that produces a decision tree theoretically provable to be identical to the centralized counterpart. It does not assume any correlation among the data attributes. DHDT requires all participating distributed sites to have an identical database schema that is known a priori, while DIDT can accommodate heterogeneous database schema with no prior, knowledge of individual database schema. Neither of these algorithms is designed for dealing with imbalanced data. Our focus is on a privacy-preserving algorithm for imbalanced data that can work in a distributed environment. Such an algorithm should use just the statistics of data from individual hospitals. Due to this characteristic of the algorithm, it belongs to the privacy-preserving data mining (PPDM) domain. Since privacy-preserving data mining is a form of Secure Multiparty Computation (SMC) [Du and Atallah 2001; Lindell et al. 2009], our algorithm qualifies as an SMC algorithm. Preserving privacy of patients is the essence of Secure Multiparty Computation.

The proposed system uses a decision model based on voting by an ensemble of classifiers. Ensemble systems [Dasarathy and Sheela 1979; Hansen and Salamon 1990] consider the decisions of multiple experts to make a final decision. In those scenarios where sufficient representative samples of one class are not available in the data, resampling with replacement can be used for drawing subsets of the insufficient class data. Each of these subsets can be combined with subsamples of the sufficient representative class to train a different classifier. Combining these classifiers creates an ensemble [Polikar 2006]. The decision of the ensemble is based on the final vote tallied by assigning weights to the decisions of individual classifiers. In our system, each individual classifier is assigned the same voting weight. We use DIDT for building each one of the experts.

DIDT makes use of the distribution of the values of attributes across classes at individual hospitals to build the global decision tree. The data structure that captures such a distribution of values of an attribute among classes is called a crosstable matrix [Caragea et al. 2004]. Suppose a given attribute u spans m values $v_i (i = 1 \text{ to } m)$ and there are n classes $c_j (j = 1 \text{ to } n)$ in the dataset. Then the (i, j)th element of the crosstable matrix corresponding to this attribute is the count of instances with class label c_j for which attribute u has value v_i. The crosstable matrix format is kept identical across all hospitals. The layout of the crosstable matrix for attribute u is [Mathew and Obradovic 2012]:

$$\begin{array}{c|cccc} & c_1 & c_2 & \cdots & c_n \\ \hline v_1 & & & & \\ v_2 & & & & \\ \vdots & & & & \\ v_m & & & & \end{array} \tag{1}$$

The sum of crosstable matrices from participating hospitals for an attribute is defined as its global crosstable matrix. The global crosstable matrix of an attribute u provides a representation for the distribution of values of u across all classes spanning the totality of data instances. The information gain for an attribute can be computed

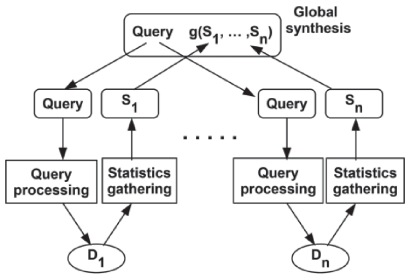

Fig. 1. Distributed model building.

using its global crosstable matrix. Assume that the global crosstable matrix for attribute u, with the layout of (1), is as follows [Mathew and Obradovic 2011]:

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

Then, the formula for the weighted average impurity measure of u is [Mathew and Obradovic 2011]:

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

Once the weighted average impurity measure for each attribute is calculated, the attribute with the smallest value of weighted average impurity measure (highest gain) is chosen for test and branching from the current node of the decision tree [Tan et al. 2006]. When a new node is constructed, the logical expression representing the path of traversal from the root to the new node is generated using Boolean operations. These logical expressions are used as seed queries for searches to globally identify the attributes for constructing the next set of crosstable matrices and eventual node split. These steps are carried out recursively to the down-levels until leaf nodes are reached.

The general computing structure of a distributed algorithm for model building involves local information gathering and global synthesis [Caragea et al. 2004]. A query generated globally is processed against the local datasets and the corresponding local statistics (partial statistics) are globally synthesized.

Figure 1 illustrates this concept. $D_1, \ldots, D_n$ are the distributed clinical data repositories, $S_1, \ldots, S_n$ are the partial statistics from each repository, and $g(S_1,..,S_n)$ represents the global synthesis. The relevance of this global synthesis by a neutral party will be evident in the next section.

3. DECISION SUPPORT SYSTEM

3.1. General Overview

The decision support system is designed to have a Clearing House (CH) that serves as the central agent to liaison with all participating hospitals. Practical reasons for a CH include vetting participating hospitals and setting up a common code of ethics that the participating hospitals are governed by. Since hospitals are careful to avoid legal ramifications and they are regulated for ethical operations, truthful information exchange

is mandatory for participating institutions. If mandates from local/state authorities or local policy of the hospital prohibit disclosure of certain information, the hospital can choose to do so. For example, certain states prohibit disclosing the HIV status of patients. Each participating hospital will have a local communication agent whose task is information exchange with the CH as well as interaction with the local database for statistics gathering. The CH is responsible for building the decision model. One of the common barriers in establishing collaborations between hospitals is the access to raw patient data. Our algorithm does not require raw patient data—only statistics of the data is needed. Hence, it is not intrusive and the likelihood of hospitals to participate is much higher. For simplicity, we assume that all sites follow the same data schema and naming convention for the attributes. The principle behind the system is that when a patient with a rare instance shows up, the medical practitioner can use the constraints around the patient's clinical attributes to dynamically build a decision model. Then the current patient attributes can be used as inputs to the decision model. The general working of the system is as follows.

A medical professional interested in making a decision based on certain attributes of interest and their associated constraints submits the query to the system. This information is passed to the CH. The CH picks a certain number of hospitals and passes along the query to the local agents. An individual agent (on behalf of the hospital) has the option to participate in the process or not. Also, hospitals with an insufficient number of instances will be eliminated from the process. After these initial negotiations, the count of hospitals participating in the process will be known. The participating agents at the individual hospitals use the query to find matching instances from the local database and respond back to the CH with matrices for the class distribution. This is a single row matrix with the number of records in each class in a predetermined order. In the discussions to follow, we refer to positive class as the class under consideration and refer to negative class as instances not in the positive class. For example, [12 3456] suggests that 12 positive class instances and 3456 negative instances meet the query criteria. If the total number of instances in a hospital is less than 4, that hospital is excluded from further participation. This is because of the possibility that statistics of small number of instances can be reverse engineered to identify patients. Also, if a hospital has no positive instances, we exclude them from further participation. These policies can be implemented at the local agent level in such a way that if the exclusion conditions are met locally, the agent can inform the CH that it is opting out. In the discussions to follow, we refer to the totality of all instances matching the query over all the hospitals after the exclusions as the *constrained data space*. The CH aggregates the individual local matrices received from the agents to estimate the balance of class distribution globally. If the ratio of positive to negative samples is below a set threshold, imbalance of class is in effect and a voting ensemble model is constructed. Otherwise, a regular distributed decision tree algorithm can be applied. The voting ensemble model is made up of distributed decision trees that are constructed using DIDT. Note that the imbalance is taken into consideration at the CH level for simplicity. It is possible that at a few individual hospital levels there is no imbalance, while there is imbalance at the CH level. An extension to this more general situation is easy to design. During the ensemble building process, the CH aggregates global crosstable matrices, calculates global information gains, selects the attribute for node splitting, keeps track of partial decision trees between iterations, and generates subqueries for down level subtrees.

The prediction model learning environment we present here is different from traditional statistical experiments where study participants are selected by some prequalification process. In classic statistical studies, samples are selected based on meeting certain qualification criteria. Our data mining algorithm makes use of opportunistic

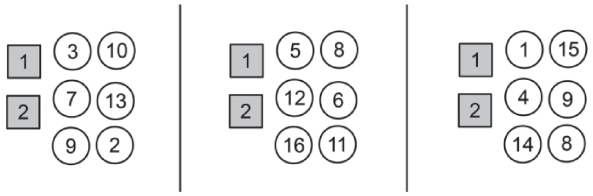

Fig. 2. A possible layout for 2 positive and 16 negative instances.

data—data that is not based on preselected study participants, but real data accumulated historically by virtue of patient visits to the hospitals. This means, the model is built based on the totality of available data instances. Hence, there is no hard rule as to how many hospitals must be involved. The algorithm by itself does not limit the number of sites to participate. The CH has the ability to select any number of sites. The goal of the decision support system is to learn from the available data so as to build a knowledge representation model. The effectiveness of the learned prediction model is rated by a success factor calculated using measures such as accuracy, Matthew's Correlation Coefficient, and so on. These measures give an indication as to how well the learning process is working.

In order to create multiple experts, at the individual participating hospital level, the negative and positive instances that are part of the constrained data space have to be processed in a specific way. In each hospital, groups of multiples of $(m + 1)$ instances are made in such a way that for 1 positive instance, m negative instances are selected. To simplify presentation in this article and the experiments, we will use $m = 3$. These groupings are repeated in a cyclic fashion. More formally, let there be k participating hospitals with p_i positives and n_i negatives that satisfy the query in the ith hospital. Let P^i denote the corresponding set of positives and N^i the related set of negatives. Let $N^i_1, N^i_2, N^i_3, \ldots$. be a random selection of $3p_i$ negative instances without replacement from N^i. We combine P^i with each of the selected subsets of N^i. If there are less than $3p_i$ instances for the last selection, then randomly select additional instances from N^i to have $3p_i$ instances. Arranging them in order, we get the layout: $P^i \cup N^i_1, P^i \cup N^i_2, P^i \cup N^i_3, \ldots$.

For e.g., assume that there are 2 positives and 16 negatives in the constrained data space in a hospital. In this case, 6 (2×3) random negatives without replacement are to be combined with the 2 positives in successive collections. The first two sets will get 12 negatives so the third set will have only 4 negatives left. Hence, 2 negatives have to be picked randomly from the original 16 negatives so as to have 6 instances in the third set. One possible selection gives the visual layout shown in Figure 2, where squares represent instances with positive classes and circles represent instances with negative classes.

Stacking all the local layouts based on the selections from the k hospitals in order, we get the arrangement in Figure 3. In this arrangement, row i corresponds to the layout from hospital i.

The totality of instances in each column contributes to a distributed collection S_j of instances from the k hospitals. A distributed decision tree DT_j that spans S_j can be built using DIDT. These decision tree DT_j's are combined to form an ensemble of classifiers where the weight for voting is the same among the individual classifiers. One issue to be addressed is the number of classifiers to be constructed. The number of classifiers that constitute the ensemble are built incrementally starting with 1 and adding 2 at a time. At each stage, the Balanced Accuracy is calculated for the ensemble

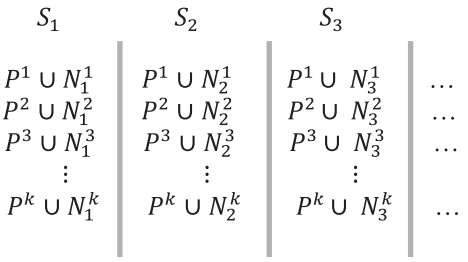

Fig. 3. Layout of the repeated positives and random negatives from all hospitals.

of all classifiers thus far. That is, Balanced Accuracy is calculated for ensembles of 1, 3, 5, 7, ... of classifiers; where the ensemble at any stage is made up of the classifiers from the previous stage plus 2 new classifiers. There are two scenarios with the Balanced Accuracies. First one is where a maximum Balanced Accuracy is reached at some stage and the immediately following ensembles have lower Balanced Accuracies. In the second case, the Balanced Accuracies increase and tend to converge to a limit.

In the first case, after the local maximum Balanced Accuracy is reached, the next 2 sets of ensembles do not produce better Balanced Accuracies. That is, we observe a sequence of 3 ensembles with the first one having maximum Balanced Accuracy. In the case of converging Balanced Accuracies, a sequence of 3 consecutive ensembles will have Balanced Accuracies very close to each other; i.e. they differ from each other by a small delta value. These two cases can be accommodated in a generalized testing condition for the algorithm to stop as follows.

Let Δ be a predefined small value. When the maximum Balanced Accuracy MaxBA is observed at an ensemble MaxENS and the Balanced Accuracies at the next 2 ensembles are less than MaxBA+Δ, the algorithm stops selecting MaxENS.

3.2. Formal Description

The complete processing from start of the query to voting ensemble assembly is done in 2 phases. In the first phase, initializations are done by the CH. In the second, the ensemble construction takes place.

Phase 1. Starts when a new query Q arrives at the CH

01. CH identify hospitals to participate
02. CH send participation request and query Q to selected hospital agents
03. Complete opt out process
04. CH collects initial class distribution matrices from all participating agents
05. Eliminate hospitals with insufficient data; Finalize list of participating hospitals
06. CH estimates data imbalance based on a preset threshold
07. If imbalance condition is detected,
    proceed with step 08
    else
    use a regular distributed classification process and finish
08. Agents do random $3p_i$ negative selections and complete the layout
09. Agents confirm readiness to CH

Phase 2. Starts after confirming readiness from all participating agents

10. CH initiates DIDT using Q as the seed query to build 1 expert
11. The Balanced Accuracy is computed using the ensemble of 1 expert
12. Denote this expert as MaxENS and the Balanced Accuracy as MaxBA

13. MaxENS→next = MaxENS→next→next = null;
14. Initialize Δ
14. While (more experts to be built) {
    if (MaxENS→next is null) {
        CH initiates DIDT using Q as seed query to build 1 more expert
        CH adds the expert to the ensemble. Denote this as MaxENS→next and
            Balanced Accuracy as MaxENS→nextBA;
    }
    if (MaxENS→next→next is null) {
        CH initiates DIDT using Q as seed query to build 1 more expert
        CH adds the expert to the ensemble
        Denote this as MaxENS→next→next and
            Balanced Accuracy as MaxENS→next→nextBA
    }
    if ((MaxENS→nextBA <= MaxBA+Δ)
            and (MaxENS→next→nextBA <= MaxBA+Δ)) {
        no more experts to build; MaxENS is the final ensemble;
    }
    else {
        if ((MaxENS→nextBA > MaxBA) and
                (MaxENS→next→nextBA > MaxENS→nextBA)) {
            MaxENS = MaxENS→next→next;
            MaxBA = MaxENS→next→nextBA;
            MaxENS→nextBA = MaxENS→next→nextBA = 0;
            MaxENS→next = MaxENS→next→next = null;
        }
        else if (MaxENS→nextBA < MaxBA) {
            MaxENS = MaxENS→next;
            MaxENS→next = null
            MaxENS→next→next = null;
            MaxBA = MaxENS→nextBA;
            MaxENS→nextBA = MaxENS→next→nextBA = 0;
        }
    }
End while
15. End of algorithm

4. COMPUTATIONAL COMPLEXITY AND COMMUNICATION COSTS

The complexity of the ensemble is dominated by the computational needs of DIDT and expressed differently for the global level and local level. Globally, it is influenced by the height of the tree H and the number of hospitals M. Locally it is influenced by the height of the tree H and the number of instances at the hospital N. The height of the tree H is bounded by the number of attributes A and in practice is usually much smaller than A as will be demonstrated in the experimental results section. Using A, M, and N:

Computational complexity at the Clearing House level is $O(A^2M)$.
Computational complexity for a member hospital is $O(A^2N)$.

The communication costs for each hospital is $O(A^2)$.

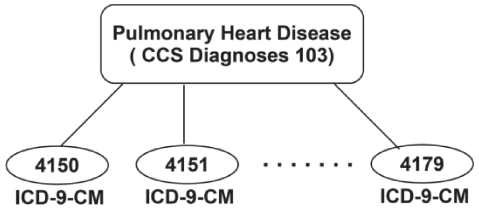

Fig. 4. Parent-child relationship between CCS code 103 and ICD-9-CM codes.

Table I. Distribution of Patient Records Based On Age

Age	Number of patients	Age	Number of patients	Age	Number of patients
0 – 7	1087392	40 – 47	585964	80 – 87	766250
8 – 15	105890	48 – 55	748453	88 – 95	311699
16 – 23	480465	56 – 63	817277	96 – 103	32537
24 – 31	649395	64 – 71	805034	104 – 111	665
32 – 39	577986	72 – 79	841738	112 – 119	17

5. DATA AND PREPROCESSING

5.1. NIS 2009 Data

For experimental evaluation, we used the Nationwide Inpatient Sample (NIS) database for 2009. The NIS 2009 database was created by 20% stratified sampling from discharge level information from all inpatients from hospitals across the USA. It was created by the Agency for Healthcare Research and Quality (AHRQ)[1] Healthcare Cost and Utilization Project (HCUP) and the database has 7,810,762 patient records from 1050 hospitals. Each data instance corresponds to an "inpatient stay record." Each patient record holds up to 25 high-level codes for diseases. These high level codes are based on HCUP Clinical Classification Software (CCS) and are developed by combining ICD-9-CM codes in a hierarchical fashion. The CCS for ICD-9-CM is a diagnosis and procedure categorization scheme [CCS 2012] in which closely related ICD-9-CM codes are combined under the same parent CCS code. For e.g. the CCS code for pulmonary heart disease is 103. The parent-child relationship between CCS diagnosis code 103 and its sibling ICD-9-CM codes is shown in Figure 4.

Only 3 of the 14 ICD-9-CM codes associated with CCS Diagnoses 103 are shown in the figure. The complete sibling ICD-9-CM codes are: 4150, 4151, 41512, 41513, 41519, 4160, 4161, 4162, 4168, 4169, 4170, 4171, 4178, and 4179. There are 259 CCS codes in total.

The distribution of patients among NIS 2009 data based on age is given in Table I.

Male patients account for 58.08% of the records and female patients account for 41.92% of the records. The patient records contain the HCUP race codes. The distribution of patient records based on HCUP race codes is given in Table II.

5.2. Preprocessing

We used the load program for SPSS, available on the AHRQ-HCUP Web site, for loading the data into SPSS Statistics software Version 19 from IBM. The data instances were exported as comma separated value (csv) files. PERL scripts were written for parsing these files to our specific needs. Software implementation for the experiments was done using JAVA.

For the purpose of our experiments, the 259 CCS codes were represented as binary attributes. The NIS 2009 database contained up to 25 CCS codes per hospitalization

[1]http://ahrq.gov

Table II. Distribution of Patient Records Based on HCUP Race Code

HCUP race code	Description	Number of patients
1	White	4358125
2	Black	909981
3	Hispanic	849907
4	Asian or Pacific Islander	176129
5	Native American	52781
6	Other	267670

record. For a given hospitalization record, the presence of a CCS code was marked by setting the value of the corresponding binary attribute as 1 and the absence of a CCS code was marked by setting the value of the corresponding binary attribute to 0. Thus, the 259 binary attributes represent the presence or absence of the corresponding CCS codes in a hospitalization record. In our experiments, we were interested in 262 attributes for each hospitalization record. They were: Age, Race, Sex, and the 259 binary attributes for CCS codes. The selection of these attributes was influenced by the work of Khalilia et al. [2011]. Values for the attribute "race" were missing from 4 states: Minnesota, North Carolina, Ohio, and West Virginia. Hence, hospitals in these states were excluded, resulting in the exclusion of 900,578 instances. Also, in some hospitals from other states, the attribute values for "race" were missing from 295,591 records. We included only data instances for patient records that had all the attributes present. Age attribute was categorized using a binning process. A range of 8 years (starting with ages 0–7) was used per bin.

6. EXPERIMENTS AND RESULTS

6.1. Experiments

In order to compare the accuracy of results, we use the concept of balanced accuracy, where $balanced\ accuracy = \frac{1}{2}(sensitivity + specificity)$.

In the following experiments, imbalance threshold was set at 2.50%. Tenfold crossvalidations were used in all experiments. For keeping crossvalidations to 10, a suitable number of hospitals were randomly selected and logically grouped into bands. A leave-one-band-out method was applied for 10-fold crossvalidations. The baseline experiments were done using DIDT, which being a Privacy Preserving Data Mining algorithm qualifies as a Secure Multiparty Computation (SMC) algorithm. Thus, we are comparing results of our ensemble SMC algorithm to the baseline SMC algorithm, DIDT. For consistency in comparing baseline DIDT and ensemble model results, the random selection of hospitals, bands, and testing sets for 10-fold crossvalidations were kept the same in both cases for all experiments.

The attributes of interest and constraints were used to create a seed query that was used to select data instances from the hospitals. The selected instances constituted the constrained data space. Two hundred and sixty-two attributes were selected for each instance; these were the 259 CCS code representations, age, sex, and ethnicity. One of the 259 CCS codes is the label used for classification. The Δ for convergence test was set at 0.1%.

6.1.1. Experiment 1. The attributes of interest were age and race. The constraints were: age between 20 and 41; race code is 5 (Native American). The problem was to classify patients in the constrained population as with or without Gastrointestinal Hemorrhage (CCS code 153).

There were 49 hospitals with positive class instances and having more than 4 instances that satisfied the query: (age $>$ 20 && age $<$ 41) && (race-code $=$ 5). In the constrained data space, there were 127 positive and 8089 negative instances. This

Table III. Summary of Balanced Accuracies

Number of Experts	Balanced Accuracy
1	0.7806
3	0.7954
5	0.7970
7	0.8254
9	0.8268
11	0.8369
13	0.8296
15	0.8376

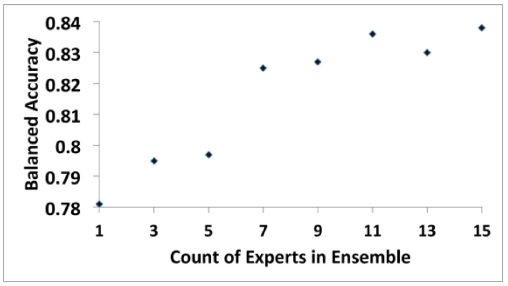

Fig. 5. Scatter plot of count of experts in the ensemble vs. balanced accuracy.

accounts for 1.55% of positives, which is less than the 2.50% threshold. Hence the imbalance condition is satisfied.

6.1.1.1. Baseline Experiment. DIDT was applied to these 49 hospitals in the constrained data space. Ten bands of hospitals were created by randomly picking 5 hospitals per band into 9 bands and the remaining 4 into the 10^{th} band. Leave-one-band-out 10-fold crossvalidations using DIDT, resulted in a balanced average accuracy of 59.39%.

6.1.1.2. Ensemble Method Experiment. For fair comparison with the baseline experiment, the same random order of hospitals was used. We did tenfold crossvalidations, using the ensembles of classifiers per crossvalidation. Corresponding to each training set in the baseline experiment, experts were built using DIDT per crossvalidation. Testing sets for each crossvalidation were kept the same as the ones from the baseline experiment.

Starting with 1 expert and adding 2 experts at a time, we got the results shown in Table III. The plot of balanced accuracies for 1, 3, 5, ... experts is shown in Figure 5.

As seen from Table III, the ensemble with 11 experts had maximum balanced accuracy so far. Also, 13 and 15 experts did not produce balanced accuracies more than $0.8369 + \Delta = 0.8369 + 0.1\%$ (0.8379). Hence, the ensemble with 11 experts and balanced accuracy 83.69% was chosen.

Average depth of the trees among the 11 ensembles was 35, while the number of independent attributes was 261. The average number of attributes that were used for node splits was 52. The most influential attribute in the decision trees was CCS code 196; normal pregnancy and/or delivery.

6.1.2. Experiment 2. The attributes of interest were age and state. The constraints were: age between 39 and 51; hospital state is one of GA, SC, and VA.

The problem was to classify patients in the constrained population as with or without peripheral and visceral atherosclerosis (CCS code 114).

Table IV. Summary of Balanced Accuracies

Number of Experts	Balanced Accuracies
1	0.6370
11	0.6773
13	0.6816
15	0.6781
17	0.6803

Table V. Summary of Balanced Accuracies

Number of Experts	Balanced Accuracies
25	0.8064
27	0.8067
29	0.8057

Sixty-four hospitals satisfied the participation criteria and the query: (age $>$ 39 && age $<$ 51) && (hospital-state in {'GA','SC','VA'}). In the constrained data space, there were 1259 positive and 60,272 negative instances. This accounts for 2.05% of positives, which is less than the 2.50% threshold and hence the imbalance condition is in effect.

6.1.2.1. Baseline Experiment. 7 hospitals were randomly picked per band from 4 bands and the remaining 36 hospitals were grouped into 6 bands of 6 randomly selected hospitals. DIDT using tenfold crossvalidations by leave-one-band-out method resulted in a balanced average accuracy of 54.48%.

6.1.2.2. Ensemble Method Experiment. Keeping the same random order of hospitals and training/testing sets as the baseline experiment, 10-fold crossvalidations with ensembles of 1,3,5,... resulted in the balanced accuracies shown in Table IV.

As seen from Table IV, the ensemble with 13 experts had maximum balanced accuracy so far and the ensembles with 15 and 17 experts did not produce balanced accuracies better than $0.6816 + \Delta$ (i.e. 0.6826). Thus the final selection was the ensemble with 13 experts and a balanced accuracy of 68.16%.

Average depth of the decision trees among the ensembles was 59 compared to the number of independent features, 261. On average 162 attributes were used for node splits. The most influential independent attribute was CCS Code 101; coronary atherosclerosis and other heart disease.

6.1.3. Experiment 3. The attributes of interest were age and state. The constraints were: age between 27 and 40; hospital state is CA. The problem was to classify patients in the constrained population as with or without Pancreatic disorders (not diabetes)—CCS code 152.

There were 75 qualifying hospitals that satisfied the query: (age $>$ 27 && age $<$ 40) && (hospital-state is 'CA'). 2190 positive and 108052 negative instances in the constrained data space accounts for 1.99% of positives, which is less than the 2.50% threshold, resulting in an imbalance condition.

6.1.3.1 Baseline Experiment. 10 bands of hospitals were created by randomly picking 8 hospitals per band for the first 5 bands and 7 hospitals per band for a second set of 5 bands. Tenfold cross-validations by the leave-one-band-out method using DIDT resulted in a balanced average accuracy of 56.18%.

6.1.3.2 Ensemble Method Experiment. Keeping the same random order of hospitals and training/testing sets as the baseline experiment, tenfold crossvalidations with ensembles of 1,3,5,... resulted in closing balanced accuracies shown in Table V.

Table VI. Summary of Prediction Statistics

Experiment	Balanced prediction accuracy			Matthew's Correlation Co-efficient	
	Baseline	Ensemble	Improvement	Baseline	Ensemble
1	59.39%	83.69%	24.30%	0.17	0.22
2	54.48%	68.16%	13.68%	0.08	0.12
3	56.18%	80.64%	24.46%	0.12	0.24

Table VII. Summary of Prediction Statistics

Experiment	Sensitivity		Specificity	
	Baseline	Ensemble	Baseline	Ensemble
1	0.20	0.84	0.98	0.83
2	0.11	0.60	0.97	0.77
3	0.14	0.75	0.98	0.87

As seen from Table V, the ensemble with 25 experts had the maximum balanced accuracy so far and the ensembles with 27 and 29 experts did not produce balanced accuracies better than $0.8064 + \Delta$ (0.8074). Thus, the algorithm stops with the selection of 25 experts and balanced accuracy 80.64%.

Average depth of the trees among ensembles was 107 which is much smaller than the number of independent attributes, 261. The average number of attributes used for node splits was 193. The most influential attribute was CCS Code 196; normal pregnancy and/or delivery.

6.2. Comparison of Results

As can be surmised from Tables VI and VII, the ensemble-based experiments gave improvements in balanced prediction accuracies ranging from 13.68% to as much as 24.46% over the baseline experiments. These significant improvements were achieved by preserving data privacy. MCC also showed improvements. By repeated grouping of positive instances with negative instances in a ratio of 1:3 and using the decisions of multiple experts, the errors in classification by a single classifier in the baseline is improved substantially by smoothing over multiple experts' opinions.

7. CONCLUSIONS

In this article, we developed a system for distributed clinical decision support when the data instances have class imbalance issues. The system can dynamically detect the class imbalance and build the classifier model without the need for actual data from the participating hospitals. The decision model we built for the system is a voting machine that is an ensemble of decision trees built individually in a distributed manner using the DIDT algorithm. The number of experts for the ensemble can be estimated by the system. Each member of the ensemble has the same voting weight. Since only statistics of the data and no real data are used in the process, patient privacy is preserved. The model-building makes use of oversampling of insufficient class data. We illustrated the working of the system using the NIS 2009 dataset and showed that the system improves the balanced accuracy and Matthew's Correlation Coefficient over the baseline system.

ACKNOWLEDGMENTS

Nationwide Inpatient Sample (NIS), Healthcare Cost and Utilization Project (HCUP), Agency for Healthcare Research and Quality, provided data used in this study.

REFERENCES

Arling, P. A., Doebbeling, B. N., and Fox, R. L. 2011. Leveraging social network analysis to improve implementation of evidence-based practices and systems in healthcare. In *Proceedings of the 44th Hawaii International Conference on System Sciences.* 1–10.

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

Bobrow, D. G., Mittal, S., and Stefik, M. J. 1986. Expert systems: Perils and promise. *Comm. ACM, 29*, 9, 880–894.

Buchanan, B. G. and Shortliffe, E. W. 1984. *Rule Based Expert Systems: The MYCIN Experiments in the Stanford Heuristic Programming Project,* Addison-Wesley, Reading, MA.

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

CCS. Clinical Classifications Software (CCS) for ICD9CM. Appendix A: SingleLevel Diagnoses. http://www.hcupus.ahrq.gov/toolssoftware/ccs/ccs.jsp.

Dasarathy, B. V. and Sheela, B. V. 1979. Composite classifier system design: Concepts and methodology. *Proc. IEEE. 67*, 5, 708–713.

Dolin, R. H., Alschuler, L., Beebe, C., Biron, P. V., Boyer, S. L., Essin, D., Kimber, E., Lincoln, T., and Mattison, J. E. 2001. The HL7 clinical document architecture. *J. Amer. Med. Inform. Assoc. 8*, 552–569.

Du, W. and Atallah, M. J. 2001. Secure multiparty computation problems and their applications: A review and open problems. In *Proceedings of the New Security Paradigms Workshop*. 11–20.

Ertekin, S., Huang, J., Bottou, L, and Giles, C. L. 2007. Learning on the border: Active learning in imbalanced data classification. In *Proceedings of CIKM.*

Hansen, L. K. and Salamon, P. 1990. Neural network ensembles. *IEEE Trans. Pattern Anal. Mach. Intell. 12*, 10, 993–1001.

Japkowicz, N. 2000. The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence.* 111–117.

Jin, R. and Agrawal, G. 2003. Communication and memory efficient parallel decision tree construction. In *Proceedings of the 3rd SIAM International Conference on Data Mining (SDM).* 119–129.

Kansagara, D., Englander, H., Salanitro, A., Kagen, D., Theobald, C., Freeman, M., and Kripalini, S. 2011. Risk prediction model for hospital readmission. *J. Amer. Med. Assoc. 306*, 15, 1688–1698.

Khalilia, M., Chakraborty, S., and Popescu, M. 2011. Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Inform. Decis. Making, 11*, 51.

Lindell, Y. and Pinkas, B. 2009. Secure multiparty computation for privacy-preserving data mining. *J. Priv. Confident. 1*, 1, 59–98.

Mathew, G. and 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.* 129–134.

Mathew, G. and Obradovic, Z. 2012. Distributed privacy preserving decision system for predicting hospitalization risks in hospitals with insufficient data. In *Proceedings of the 11th International Conference on Machine Learning and Applications.* 178–183.

Moret, B. M. E. 1982. Decision trees and diagrams. *ACM Comput. Surv. 14*, 4, 593–623.

Oster, S., Langella, S., Hastings, S., Ervin, D., Madduri, R., Kurc, T., Siebenlist, F., Covitz, P., Shanbhag, K., Foster, I., and Saltz, J. 2007. In *Proceedings of AMIA Annual Symposium.* 573–577.

Polikar, R. 2006. Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* Third Quarter, 21–45.

Popescu, M. and Khalilia, M. 2011. Improving disease prediction using ICD9 ontological features. In *Proceedings of the IEEE International Conference on Fuzzy Systems*.

Quinlan, J. R. 1986. Induction of decision trees. *Mach. Learn. 1,* 81–106.

Risch, N. J. 2000. Searching for genetic determinants in the new millennium. *Nature 405*, 847–856.

Sim, I., Gorman, P., Greense, A., Haunes, R. B., Kaplan, B., Lehman, H., and Tang, P. C. 2001. Clinical decision support systems for the practice of evidencebased medicine. *J. Amer. Med. Inform. Assoc. 8*, 6, 527–534.

Sittig, D. F., Krall, M. A., Dykstra, R. H., Russell, A., and Chin, H. L. 2006. A survey of factors affecting clinician acceptance of clinical decision support system. *BMC Med. Inform. Decis. Making 6*, 6.

Stiglic, G., Pernek, I., Kokol, P., and Obradovic, Z. 2012. Disease prediction based on prior knowledge. In *Proceedings of the ACM SIGKDD Workshop on Health Informatics, in Conjunction with 18th SIGKDD Conference on Knowledge Discovery and Data Mining*.

Tan, P., Steinbach, M., and Kumar, V. 2006. *Introduction to Data Mining.* Pearson Addison Wesley Boston, MA, 160.

Warren, R., Solomonides, A. E., Del Frate, C., Warsi, I., Ding, J., Odeh, M., McClatchey, R., Tromans, C., Brady, M., Highnam, R., Cordell, M., Estrella, F., Bazzochi, M., and Amendolia, S. R. 2007. MammoGrid – A prototype distributed mammographic database for Europe. *Clinical Radiol. 62*, 1044–1051.

Wier, L. M., Elixhauser, A., Pfunter, A., and Au, D. H. 2011. Overview of hospitalizations among patients with COPD, 2008; Statistical Brief #106. http://www.hcupus.ahrq.gov/reports/statbriefs/sb106.jsp.

Yu, W., Liu, T., Valdez, R., Gwinn, M., and Khoury, M. J. 2010. Application of support vector machine modeling for prediction of common diseases: The case of diabetes and prediabetes. *BMC Med. Inform. Decis. Making,* 10, 16.

Received November 2012; revised June 2013; accepted August 2013