Distributed Privacy Preserving Decision System for Predicting Hospitalization Risk in Hospitals with Insufficient Data

George Mathew, Zoran Obradovic
Center for Data Analytics and Biomedical Informatics
Temple University
Philadelphia, PA
{George.Mathew,Zoran.Obradovic}@temple.edu

Abstract— **Building prediction models for suggestive knowledge from multiple sources dynamically is of great interest from a clinical decision support point of view. This is valuable in situations where the local clinical data repository does not have sufficient number of records to draw conclusions from. However, due to privacy concerns, hospitals are reluctant to divulge patient records. Consequently, a distributed model building mechanism that can use just the statistics from multiple hospitals' databases is valuable. Our DIDT algorithm builds a model in that fashion. In this study, using National Inpatient Sample (NIS) data for 2009, we demonstrate that DIDT algorithm can be used to help collaboratively build a better decision-making model in situations where hospitals have small number of records that are insufficient to make good local models. Based on 262 attributes used for model building, we showed that 9 collaborating hospitals each with less than 100 cases of hospitalizations related to diabetes were able to achieve 9.9% improvement in accuracies of hospitalization prediction collectively using a distributed model as compared to relying on local models developed on their own. When relying on local risk prediction models for diabetes at these 9 hospitals, 159 of 357 patients were misclassified and prediction was impossible for another 16 patients. Our integrated model reduced the misclassification to 138 effectively providing accurate early diagnostics to 37 additional patients. We also introduce the concept of banding to improve DIDT algorithm so as to logically combine multiple hospitals when large number of hospitals is involved for reduction in cross-validation folds.**

Keywords- distributed decision making; privacy preserving prediction model; hospitalization risk prediction

I. INTRODUCTION

Applying data mining techniques to clinical domain data can help with decision support systems and in identifying at risk patients for targeted communications [1]. Since medicine is "characterized by much judgmental knowledge" [2], suggestions for decision-making are valuable to a practitioner. Other prediction models of recent interest are hospital readmission cost [3] and health insurance underwriting [4]. In real life scenarios, the databases from various hospitals are distributed geographically. There is interest in building decision support systems that can harness the power of collective intelligence from multiple hospitals using the power of Internet [5]. Survey results have shown that physicians are interested in such decision support systems [6]. Collecting data from all the distributed hospitals to a central location is not practical due to privacy concerns and regulatory implications. Hence, a distributed model building mechanism is attractive. Since hospitals are reluctant to divulge patient records to other institutions due to legal and compliance issues, algorithms that can build prediction models in a distributed environment using just the statistics of the data are very useful. Furthermore, local databases in some hospitals do not have sufficient number of records of a certain diagnosis to garner intelligence from. In these cases, mining the collective distributed space of similar hospitals in a collaborative fashion can possibly lead to a quite useful decision making model. For e.g., 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. The objective of our study is to help draw conclusions on a certain diagnosis when local samples are insufficient, using shared statistics from multiple hospitals. 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. Based on this premise, we identified nine hospitals in the NIS (Nationwide Inpatient Sample) 2009 data set, each of which had less than 100 patient records having diabetes mellitus without complications. For those nine hospitals we built the local models and compared them to the distributed model built using DIDT (Distributed Id3-based Decision Tree) [7] algorithm. The distributed model using just the statistics of data provided an improvement of 9.9% in accuracy over average local model accuracies. In Section II, we outline some of the salient features of the NIS 2009 data set.

DIDT is a simple algorithm that produces a decision tree identical to the one produced on an equivalent centralized data aggregation. A decision tree [8] is a data structure that represents the paths of traversals in a decision making process for classification problems. Id3 [9] is one of the commonly used decision tree building algorithms. The algorithm uses only statistics of data from the distributed hospital databases. Thus 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 e.g. if a

978-0-7695-4913-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICMLA.2012.180

IEEE computer society

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 [10] 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 [11] that introduces noise to the statistics and hence introduce distortion to the results.

II. NIS 2009 DATA

In this work, we used the Nationwide Inpatient Sample (NIS) Database for 2009 that was created by Agency for Healthcare Research and Quality (AHRQ) [12] Healthcare Cost and Utilization Project (HCUP). It contains discharge level information of all inpatients from a 20% stratified sample of hospitals across USA. The 2009 NIS database has close to eight million records from 1050 hospitals. Each data instance represents an "inpatient stay record". Due to confidentiality laws, records with some very specific medical conditions and procedures (e.g. HIV/AIDS or abortion) are not released by certain hospitals. The data records are de-identified and do not contain personally identifiable information such as name or home address. Hence, these records are the ideal subset of vertically partitioned attributes that are candidates for participation in a privacy preserving distributed decision support model. Our DIDT algorithm do not use the attributes directly, but use only statistics about these attributes. The variations in data records between hospitals provide a real world setting for studying distributed algorithms. The distribution of patient records based on age is given in Figure 1.

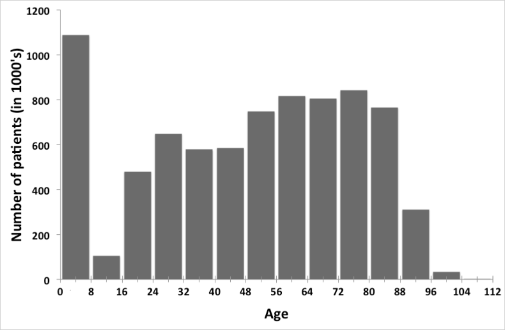

Figure 1. Distribution of patient records based on age.

There are up to 25 high level codes for diseases per data instance in the NIS 2009 data set. These are codes based on HCUP Clinical Classifications Software (CCS), developed by combining ICD-9-CM codes in a hierarchical fashion. For example, CCS code for diabetes mellitus without complications that is studied in this article is 49. The Clinical Classifications Software (CCS) for ICD-9-CM is a diagnosis and procedure categorization scheme [13] where closely related ICD-9-CM codes are combined under a parent CCS code. There are a total of 259 CCS codes in all. The largest percentages of records in the 2009 NIS data set were based on Essential Hypertension (CCS code 98) with 31.2% and Coronary Atherosclerosis (CCS code 101) with 31.18%. Our study was focused on patients with "Diabetes mellitus without complications" (CCS code 49), which accounted for 14.88% of patient records and is the fifth most common diagnosis. The parent-child relationship with CCS Diagnoses 49 and its sibling ICD-9-CM codes is shown in Figure 2.

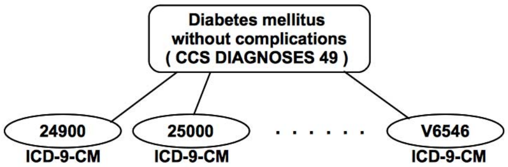

Figure 2. Parent-child relationship between CCS code 49 and ICD-9-CM codes.

Only 3 of the 12 children ICD-9-CM codes are shown in Figure 1. The complete sibling ICD-9-CM codes are: 24900, 25000, 25001, 7902, 79021, 79022, 79029, 7915, 7916, V4585, V5391, and V6546. The distributions of male and female patients were 58.08% and 41.92% respectively. The distribution of patient records based on race is given in Figure 3. The distribution of race is based on the HCUP race code.

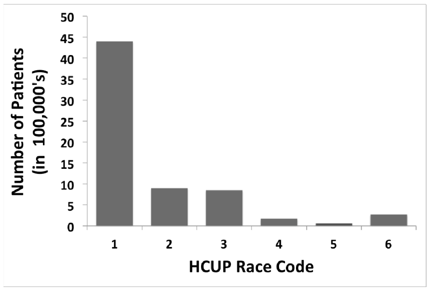

Figure 3. Distribution of patient records based on HUCP race code.

The distribution of the 5 most common specific comorbidities among the patient records in the NIS 2009 data were as given in Table 1.

TABLE I. MOST SPECIFIC COMORBIDITIES AMONG THE NIS 2009 PATIENT RECORDS

CCS Code	Description	Prevalence
98	Essential Hypertension	31.20%
101	Coronary Atherosclerosis	31.18%
106	Cardiac Dysrhythmias	16.80%
108	Congestive heart failure	15.14%
49	Diabetes mellitus without complications	14.88%

III. RELATED WORKS

The NIS data sets have been used in various medical studies with a statistical approach. Age related cholecystectomy [14] analysis was done using NIS data from 1996-2001. Factors affecting length of hospital stay in connection with mouth cellulitis [15] was done using NIS 2008 data. Hospitalization costs and post discharge follow-up care costs associated with meningococcal disease was studied [16] 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 [17]. 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 & 2009 NIS data [18]. Random Forest technique for predicting disease risks was applied by Khalilia et al. [1] 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 [19]. These data mining models were all based on centralized data architecture. In real life, the patient records are distributed among clinical databases in various hospitals. Due to this natural distribution of patient data among hospitals, a distributed data mining technique would align well with the distributed data topology. DIDT (Distributed Id3-based Decision Tree) [7] is one of the distributed decision making algorithms that builds a classification model using decision trees. In this paper, the classification problem of identifying patients as diabetic (CCS code 49) or non-diabetic, is used as the underlying basis for our study.

IV. Methodology

DIDT is a distributed decision tree building algorithm that makes use of the distribution of the values of an attribute among classes at individual hospitals. The data structure capturing this information is called a crosstable matrix [20]. Suppose an attribute u takes m values $v_1, v_2, \ldots, v_m$ and spans n classes $c_1, c_2, \ldots, c_n$ among data instances within a given hospital. Then the $(x,y)^{\text{th}}$ element of the crosstable matrix represents the number of data instances belonging to class c_y for which the attribute u has value v_x. The format of crosstable matrices are uniformly maintained among all participating hospitals. For the attribute u, as described above, the crosstable matrix takes the template form:

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

The sum of the crosstable matrices from individual hospitals, named global crosstable matrix, gives the global distribution of the values across all classes for a given attribute. The global crosstable matrices can be used to calculate the information gains and pick the attribute that gives maximum gain to decide on the branching of the decision tree. Let the global crosstable matrix for attribute u based on template (1) be as follows:

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

Then, the weighted average impurity measure for attribute u is calculated using the formula:

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

The weighted average impurity measure for each attribute is calculated using the corresponding global crosstable matrix. The attribute with the smallest value of weighted average impurity measure (highest gain) is chosen for node split [21]. At each node, the logical expression for the path from root to the node is constructed using Boolean operations. These Boolean expressions are used as search expressions so as to globally construct the data attributes to be considered for the next set of cross-table matrix evaluations and eventual node split. The process is repeated recursively till leaf nodes are reached. Cross-validation is done by leave-one-hospital-out method. In this method, data from one hospital is used for testing while data from all other hospitals are used for training.

We modified the published DIDT algorithm to accommodate a variation of crossvalidation. In the original DIDT algorithm, a leave-one-hospital-out cross-validation method was employed. When large number of hospitals are involved, this makes the number of crossvalidations quite high. To avoid this aberration, a few hospitals can be banded together to create a logical mega-hospital so that when the leave-one-hospital-out method is employed, a mega-hospital can be left out for testing. We implemented the mega-hospital building by selecting appropriate number of hospitals randomly without replacement for each mega-hospital such that these mega-hospitals provide a partition of all the hospitals. For example, if there are 250 hospitals involved, a leave-one-hospital-out crossvalidation leads to 250-fold cross-validations. However, if sets of 25 hospitals (picked randomly without replacement) are used to create 10 mega-hospitals, then only 10-fold cross-validations using the mega-hospitals need to be performed. In our study, the modified DIDT was used for the distributed model building, while local models were built using Weka opensource software [22] with 10-fold crossvalidations.

The 259 CCS codes were represented as binary attributes. The NIS data had up to 25 CCS codes per hospitalization record. For a given hospitalization record, if a CCS code was present, the value of the corresponding binary attribute was set as 1 and if a CCS code was not present, the value of the corresponding binary attribute was set to 0. Thus, the 259 binary attributes represent the presence or absence of the corresponding CCS codes in a hospitalization record. In 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 work [1]. Values for the attribute 'race' was missing from 4 states: Minnesota, North Carolina, Ohio and West Virginia. Hence, hospitals in these states were excluded. Also, in some 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. Age attribute was categorized using a binning process. A range of 8 years (starting with ages 0-7) was used for one bin. There were 1162186 diabetes patient records out of which 672683 had all 262 attributes present.

V. Experiments

A. *Pre-processing*

The NIS 2009 data was loaded into SPSS Statistics software Version 19 from IBM, using the load program supplied on the AHRQ-HUP web site. The data records were exported as comma separated values (csv) from SPSS. The csv files were parsed for creating arff formatted files using PERL scripts written for this purpose. 'arff' is one of the data input formats supported by Weka software.

We used a graph model as the underlying data framework for the distributed databases. Graph databases [23] help capture the structure of clinical data in a very natural way. When each patient record is treated as a graph, the symptoms can be represented as labeled vertices of a graph. A graph database is well suited to represent the heterogeneous patient graphs. In our experiments, the neo4j [24] opensource graph database was used for storing individual hospital data, one database per hospital, to create the distributed environment. Lucene [25] indexing was used for text indexing within the neo4j databases. Software implementation was done in JAVA.

B. *Results*

We studied the problem of classifying patients with or without "Diabetes mellitus without complications". Decision tree building on individual hospitals were done using Weka software. The experiments were based on data from hospitals with all 262 attributes present. There were 902 hospitals for which data instances existed with non-missing values for age, sex and race. The local models for these 902 hospitals were generated with 10-fold crossvalidations, resulting in the distribution of accuracy ranges shown in Table 2.

TABLE II. Distribution Of Accuracies For 902 Hospitals with Diabetes Records

Range of accuracies	Number of hospitals
could not build model	5
< 50%	1
50% - 60%	17
60% - 70%	86
70% - 80%	411
80% - 90%	353
90% - 100%	29

As seen from Table 2, there are 23 hospitals with less than 60% accuracy. To understand the distribution in this range and to identify an area of improvement, we plotted the distribution of patient records among these 23 hospitals. The result is shown in Figure 4.

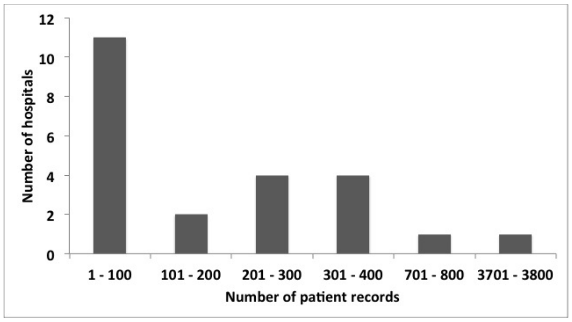

Figure 4. Distribution of hospitals with less than 60% accuracies in local models.

It is seen from this graph that the prevalence of hospitals in this group had less than 100 patient records. Hence, we decided to work on these 11 hospitals with less than 100 patient records. Of the 5 hospitals that could not build local models, 2 had less than 3 records. Since using 1 or 2 records can lead to reverse-identifying the patient(s) in a distributed system, we decided to leave out these hospitals in our study. Thus we focused on the 9 hospitals with less than 100 patient records with all 262 attributes present. We performed local decision trees building with 10-fold crossvalidations. The results are shown in Table 3.

TABLE III. Distribution Of Accuracies For Hospitals with Less Than 100 Records Having Diabetes Attribute

Number of records	Number of hospitals	Local Model Accuracy
< 10	3	-
25 - 50	2	56% - 60%
51 - 75	3	48.48% - 58.06%
76 - 100	1	57.5%

The average of local model accuracies among the 9 hospitals was calculated using:

$$\frac{\textit{count of correctly classified instances in all 9 hospitals}}{\textit{total instances in all 9 hospitals}}$$

$$= 53.08\%$$

Next, we ran the distributed DIDT algorithm on the same set of hospitals. This resulted in an accuracy of 63%, an improvement of 9.92%. To compare the results obtained by the distributed model with the equivalent centralized model, we combined all data from the 9 hospitals centrally and built decision tree using weka software. To do this centralized operation, all raw data from the 9 hospitals had to be combined. We used the same crossvalidation formats as DIDT to avoid crossvalidation mismatch. The result is shown in the last column of Table 4.

TABLE IV. COMPARISON OF ACCURACIES FOR HOSPITALS WITH LESS THAN 100 RECORDS

Number of records	Number of hospitals	Average local model accuracy	DIDT accuracy	Centralized equivalent accuracy
1 - 100	9	53.08%	63%	64.07%

It can be seen from the results that the DIDT algorithm gives empirical result close to the centralized equivalent value. The aberration in the result is due to the fact that in the tree building node split, it is possible to have multiple attribute selection choices for a node split and hence the distributed tree is not necessarily identical to the centralized one. However, the big advantage with the DIDT over the centralized equivalent tree building is that no patient data is required from the hospitals - only statistics about the patient data is required. The centralized equivalent tree building requires all patient data centrally and is costly in terms of data transfer as well as in terms of data privacy. And, in terms of not harming patients by improved diagnosis, the statistics for these methods are shown in Table 5.

TABLE V. DISTRIBUTION OF ACCURACIES FOR HOSPITALS WITH LESS THAN 100 RECORDS

Method	Average accuracy	Number of patients incorrectly diagnosed
Local models	53.08%	159*
DIDT	63%	138
Centralized equivalent	64.07%	136

*: excluding the 16 patients from 3 hospitals without local models

The results in Table 5 show another big advantage for DIDT. The number of patients incorrectly classified is quite less with DIDT compared to what the 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. Our collaborative distributed model reduced the misclassification from 159 to 138 effectively providing accurate early diagnostics to 37 additional patients.

Based on the values in Table 4, the improvement in accuracy using DIDT was 9.92%. To validate our hypothesis that the net improvement in accuracy is best for these cohorts, we computed the net improvement in accuracies when these hospitals collaborate with hospitals having higher number of similar patient records. In order to do this, DIDT algorithm was used to generate the corresponding decision trees by making use of the related patient records from the 9 hospitals plus the hospitals in the corresponding tier. The resulting improvements in accuracies for various tiers are shown in Table 6.

TABLE VI. DISTRIBUTION OF ACCURACIES FOR HOSPITALS WITH LESS THAN 100 RECORDS

Number of records	Number of hospitals	Improvement in accuracy using DIDT
1 - 100	9	9.92%
1 – 250	12	4.92%
1 - 500	19	1.49%
1 - 1000	20	0.37%

It is observed from Table 6 that the net improvement in accuracy is best when the disadvantaged hospitals with less than 100 patient records used DIDT to build a distributed prediction model. Hospitals with larger number of records do not contribute substantially to improve the accuracy when hospitals with insufficient number of data build the models collaboratively with them. Figure 5 shows a graphical representation of this trend.

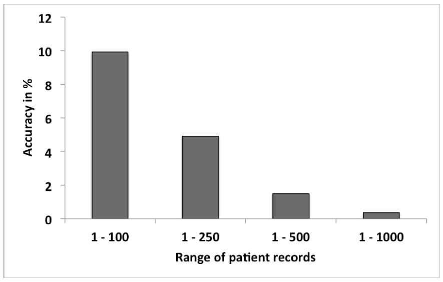

Figure 5. Distribution of accuracies among hospitals at various resolutions.

In this case, the hospitals with large number of samples to build their local models are better off by themselves as they can build a model specific to their patients.

C. *Dimension Reduction*

The next experiment was oriented towards reducing the dimension of the patient data. It was observed that some symptoms had very low frequency in the aggregated data set. Hence attributes corresponding to these symptoms with frequency less than 4 in the combined data set were eliminated. This resulted in a dimension reduction from 262 to 211 attributes. In this scenario, all 211 attributes were fully populated in all hospitals. The accuracy in this case for DIDT related results are shown in Table 7.

TABLE VII. DISTRIBUTION OF PREDICTION ACCURACIES AFTER DIMENSION REDUCTION FOR HOSPITALS WITH LESS THAN 100 DIABETES-RELATED RECORDS

Number of records	Number of hospitals	Average Local Model	DIDT accuracy	Centralized equivalent accuracy
1 - 100	9	53.08%	63%	63.40%

Comparing Tables 4 and 7, it is observed that the accuracy for DIDT remains the same even after considerable reduction in dimension.

VI. Conclusion

Using NIS data for 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 the use of DIDT. It was observed that hospitals with enough instances to create a reasonably good local model do not contribute much to improve the overall accuracy of a distributed model. 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 due to diabetes without complications. However, this methodology has no dependency on the disease per say and so can be applied to building 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 mega-hospital concept by banding together hospitals. 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

[1] M. Khalilia, S. Chakraborty, and M. Popescu. "Predicting Disease Risks From Highly Imbalanced Data Using Random Forest", BMC Medical Informatics and Decision Making 2011, http://www.biomedcentral.com/1472-6947/11/51.

[2] W. van Melle, "MYCIN: A Knowledge-based Consultation Program for Infectious Disease Diagnosis", International Journal of Man-machine Studies, Vol. 10, Issue 3, May 1978, pp. 313-322.

[3] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman, and S. Kripalani. "Risk Prediction Models for Hospital Readmission", JAMA, October 19, 2011. Vol 306, No. 15. pp. 1688 – 1698.

[4] V. Fuster et al, "Medical Underwriting for Life Insurance", McGraw-Hill's AccessMedicine 2008.

[5] D. F. Sittig, A. Wright, J. A. Osheroff, B. Middleton, J. M. Teich, J. S. Ash, E. Cambell, and D. W. Bates, "Grand Challenges in Clinical Decision Support", Journal of Biodical Informatics, Vol 41. 2008, pp. 387-392, doi:10.1016/j.jbi.2007.09.003

[6] D. F. Sittig, M. A. Krall, R. H. Dykstra, A. Russell, and H.L. Chin, "A Survey of Factors Affecting Clinican Acceptance of Clinical Decision Support", BMC Medical Informatics and Decision Making, 6:6, 2006, doi:10.1186/1472-6947-6-6

[7] G. Mathew, and Z. Obradovic, "A Privacy-Preserving Framework for Distributed Clinical Decision Support", Proceedings of the 1st IEEE International Conference on Computational Advances in Bio and medical Sciences. Feb 2011, Orlando, FL. doi:10.1109/ICCABS.2011.5729866

[8] B. M. E. Moret, "Decision Trees and Diagrams", ACM Computing Surveys, Vol 14(4), pp. 593-623, Dec 1982.

[9] J. R. Quinlan, "Induction of Decision Trees", Machine Learning, Vol. 1, pp. 81-106, March 1986.

[10] D. Caragea, A. Silvescu, and V. Honavar, "A Framework for Learning from Distributed Data Using Sufficient Statistics and its Applications to Learning Decision Trees", International Journal on Hybrid Intelligent Systems, Vol 1, Issue 1-2, April 2004, pp. 80-89.

[11] C. Dwork, "Differential Privacy", Proceedings of 33rd International Colloquium on Automata, Languages and Programming, Juy 2006, Venice, Italy, pp. 1-12.

[12] Agency for Healthcare Research and Quality. Home page: http://ahrq.gov

[13] Clinical Classifications Software (CCS) for ICD-9-CM. Appendix A: Single-Level Diagnoses. Available at: http://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

[14] S. Kuy, J. A. Sosa, S. A. Roman, R. Desai, and R. A. Rosenthal, "Age Matters: A Study of Clinical and Economic Outcomes Following Cholecystectomy in Elderly Americans", The American Journal of Surgery, Vol 201, No. 6, pp. 789-796, June 2011.

[15] M. K. Kim, R. P. Nalliah, M. K. Lee, and V. Allareddy, "Factors Associated With Length of Stay and Hospital Charges for Patients Hospitalized with Mouth Cellulitis", Oral Surgery, Oral Medicine, Oral Pathology, Oral Radiology, Vol. 113, Issue 1. Jan 2012, pp. 21-28.

[16] K. L. Davis, D. A. Misurski, J. M. Miller, T. J. Bell and B. Bapat, "Cost of Acute Hospitalization and Post-discharge Follow-up Care for Meningococcal Disease in the United States", Human Vaccines, 7:1, Jan 2011, pp. 96-101.

[17] W. Yu, T. Liu, R. Valdez, M. Gwinn and M. J. Khoury, "Application of Support Vector Machine Modeling for Prediction of Common Diseases: The Case of Diabetes Pre-diabetes", BMC Medical Informatics and Decision Making, Mar 2010, 10:16. doi:10.1186/1472-6947-10-16

[18] G. Stiglic, I. Pernek, P. Kokol, and Z. Obradovic, "Disease Prediction Based on Prior Knowledge", Proceedings of ACM SIGKDD Workshop on Health Informatics, in conjunction with 18th SIGKDD Conference on Knowledge Discovery and Data Mining, Beijing, China, August 2012.

[19] M. Popescu, and M. Khalilia, "Improving Disease Prediction Using ICD-9 Ontological Features", 2011 IEEE International Conference on Fuzzy Systems, June 27-30, 2011. Taipei, Taiwan.

[20] A. Bar-Or, D. Keren, A. Schuster, and R. Wolff, "Hierarchical Decision Tree Induction in Distributed Genomic Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 17 Issue 8, pp. 1138-1151, Aug. 2005.

[21] P. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining", Boston, MA, USA: Pearson Addison Wesley, 2006, pp. 160.

[22] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutermann, and I. H. Witten, "The WEKA Data Mining Software: an Update", SIGKDD Explorations, Vol. 11(1), pp. 10-18, Jun. 2009.

[23] D. J. Cook, and L. B. Holder, "Mining Graph Data", Wiley Interscience, Hoboken, NJ, USA, 2007.

[24] Home page for neo4j graph database. http://neo4j.org

[25] Lucene project from Apache foundation. http://lucene.apache.org