Poster: Auto-reduction of Features for Containing Communication Costs in a Distributed Privacy-Preserving Clinical Decision Support System

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

***Abstract*—The Distributed ID3-based Decision Tree (DIDT) algorithm provides a basis for Distributed Privacy-preserving Clinical Decision Support Systems. Due to large number of features associated with clinical patient records and iterative nature of distributed algorithms, exchanging information related to all features is expensive. We show that auto-reduction for features can be achieved with significant improvement in communication costs. Auto-reduction was implemented in DIDT and results of experiments using Nationwide Inpatient Sample data sets for 2008 are presented.**

Keywords - medical informatics; feature reduction; clinical decision support systems.

I. INTRODUCTION

Distributed ID3-based Decision Tree (DIDT) [1] is a privacy-preserving algorithm that can dynamically learn from patient data spread across multiple hospitals using just statistics of data. In a distributed tree building process, the paths of traversal requires the elimination of one attribute at a time for node splitting. In each iteration, there are data communications between all hospitals. If the number of features can be reduced from root node onwards, the communication costs can be contained. In DIDT, crosstable matrices at root level of the decision tree help identify attributes that are weak or non-contributors to the tree and provide an opportunity to reduce communication costs.

II. METHODOLOGY, EXPERIMENTS & RESULTS

Consider an attribute *u* that takes a constant value *k* in all instances. The information gain at any level for *u* is zero. So, attribute *u* will not be selected for node split at the root or middle node levels and it falls through until the instances are exhausted or reaches a leaf level. In both cases, attribute *u* has no effect on the decision tree. A patient visit record will not have all the diagnosis codes. Consequently, a number of diagnoses will be sparse in a patient record. The non-sparse attributes (reflecting absence of diagnoses) will take a constant manifest value or near constant value among all instances. Eliminating attributes that have constant value in all instances, at the beginning of decision tree building process will contain communication costs.

Experiments were performed using Nationwide Inpatient Sample 2008 data that contains discharge level information of inpatients from approximately 20% stratified sample of community hospitals from 42 states in USA.

Experiment was done based on classification for a Californian teenage patient to be having "essential hypertension" (CCS Code 98) or not. Auto-elimination resulted in reduction of 26 attributes at a savings of 11.20% in communication costs with no loss of accuracy. To explore elimination of attributes with near constant values, we set threshold for the positive instances to be less than 0.0125%, 0.025% and 0.050%. Results are in Table 1.

TABLE I. COMPRISON OF RUNTIMES FOR SMALL THRESHOLDS

Number of Attributes	Threshold	Improvement in communication	Accuracy
259		n/a	96.36%
233	0.00%	11.20%	96.36%
225	0.0125%	13.41%	96.34%
220	0.0025%	16.54%	96.34%
208	0.050%	21.77%	96.34%

As seen from Table 1, setting small thresholds upto 0.050% had negligible effect on accuracy with better savings in communication costs. The effect of higher thresholds values from 1% to 5% is shown in Figure 1.

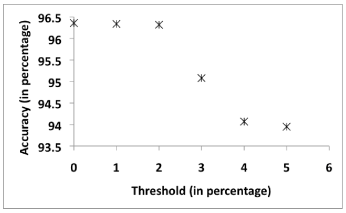

Figure 1. Scatterplot of threshold vs accurcay

ACKNOWLEDGMENT

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

REFERENCES

[1] G. Mathew, and Z. Obraodovic, "A Privacy-Preserving Framework for Distributed Clinical Decision Support", Proceedings of 1st IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS) Feb, 2011. Orlando, FL. pp. 129-134.

978-1-4673-1321-6/12/$31.00 ©2013 IEEE