A Generative Framework for Predictive Modeling using Variably Aggregated, Multi-source Healthcare Data

Yubin Park
Department of Electrical and Computer Engg.
University of Texas at Austin
Austin, TX, USA
yubin.park@utexas.edu

Joydeep Ghosh
Department of Electrical and Computer Engg.
University of Texas at Austin
Austin, TX, USA
ghosh@ece.utexas.edu

ABSTRACT

Many measures of healthcare delivery or quality are not publicly available at the individual patient or hospital level largely due to privacy restrictions, legal issues or reporting norms. Instead, such measures are provided at a higher or more aggregated level, such as state-level, county-level summaries or averages over health zones (HRR[1]s and HSA[2]s). Such levels constitute partitionings of the underlying individual level data into segments that may not match the data clusters that would have been obtained if one analyzed individual-level data. Moreover, different data sources may use different underlying partitions as the bases for their data summarization. How can one run data mining procedures such as clustering or regression on data where different variables are available at different levels of aggregation or granularity? We first examine this problem in a clustering setting given a mix of individual-level and (arbitrarily) aggregated level data. For this setting, we present an extension of the Latent Dirichlet Allocation model that can use such aggregated information. The model provides reasonable cluster centroids under certain conditions, and is extended to impute masked features at the individual-level. The imputed feature values are based on an underlying mixture distribution, and help to improve the performance in subsequent predictive modeling tasks. The model parameters are learned using an approximated Gibbs sampling method, which employs the Metropolis-Hastings algorithm efficiently. Experimental results using data from the Dartmouth Health Atlas, CDC, and the U.S. Census Bureau are provided to illustrate the generality and capabilities of the proposed framework.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Probabilistic algorithms; J.3 [**Life and Medical Sciences**]: Health

General Terms

Algorithms

Keywords

Clustering, Privacy Preserving Data Mining, Dartmouth Health Atlas, LDA, Multi-source Health Metrics

[1]Hospital Referral Region
[2]Hospital Service Area

KDD-DMH '11, August 21, 2011, San Diego, California, USA.

1. INTRODUCTION

Despite the tremendous information explosion and availability of public-domain medical and healthcare data recently (for example see www.data.gov/health), many of the health-related features or indicators are only available at a highly aggregated level, due to privacy concerns, reporting norms or legal issues [14]. In particular, routinely collected administrative data sets, such as national registers, aim to collect information on a limited number of variables for the whole population, while survey and cohort studies contain more detailed data from a sample of the population [12]. Even if the individual records are available, some features may be suppressed to protect identities of data holders. For example, Texas Department of State Health Services provides 'Texas Inpatient Public Use Data File (PUDF)', which contains data on discharges from Texas hospitals [1], but the ZIP code information in PUDF is suppressed or eliminated depending on the number of patients in a given region. As data mining algorithms should ideally be applied to individual-level data to discover valuable information, limited access to the raw entries introduces conflict of interests between data miners, patients and providers [15]. Several privacy preserving data mining algorithms have been suggested to overcome this conflict [3, 9]. However, requirements of privacy preservation are difficult to achieve for several types of analyses, and these algorithms are typically more complex and less capable compared to privacy-agnostic techniques.

Many health or healthcare indicators are available at different aggregated levels, rather than providing an entry for each individual. For example, average income by state, average death ratio by city, or average smoking rate by country are available through a variety of easily accessible public reports. Although these aggregated statistics cannot reconstruct the underlying individual-level data, these aggregated data can be combined with individual data to produce more informative models. In epidemiology, it has been observed that ecological bias from aggregate administrative data can be alleviated by incorporating surveys of individual exposures or case-control data, leading to recent

attempts at integrating data at multiple levels of summarization. In [11], *hierarchical related regression* (HRR) is introduced, which combines both aggregated and individual level datasets. The proposed HRR model alleviates ecological biases on its group-level data while improving individual-level predictabilities. However, the features in both aggregated and individual datasets need to be the same, and the generative process of the data is not considered. On the other hand, in [12], two datasets with different features are used to build a better regression model within a Bayesian framework. But both datasets need to be at the same level of aggregation.

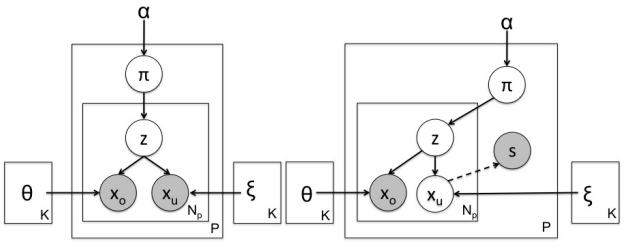
In this paper, we seek a better utilization of such aggregated information for augmenting the individual-level data. Assuming that the dataset of interest is generated by a mixture model, and that the partitions that form aggregation units (such as states or counties) contain different ratios of the mixture components, we introduce an extension of the Latent Dirichlet Allocation (LDA) model [5], using the Central Limit Theorem to capture the underlying clusters. Despite the limited nature of given aggregated information, our clustering algorithm provides not only reasonable cluster centroids, but also imputes the unobserved individual features. These imputed features reflect the underlying distribution of the data, thus a predictive model using these extended information shows improved accuracy. As many datasets in the healthcare domain are divided into multiple tables containing different levels of aggregation (sometimes obtained from different sources), the suggested methodology in this paper can be useful in maximizing the use of such available information. Our approach can easily be extended to situations where different features are aggregated over different partitions of the raw data records.

2. CLUSTERING MODEL

We denote the set of features that are available at the individual level, where "individual" refers to entities at the highest resolution available, by $\vec{x}_o$. The features that are observed only at an aggregated level are denoted by $\vec{x}_u$, where u denotes 'unobserved' at the individual level. Thus there is an underlying "complete" dataset ($\mathcal{D}_x = \{(\vec{x}_o, \vec{x}_u)_1, (\vec{x}_o, \vec{x}_u)_2, ..., (\vec{x}_o, \vec{x}_u)_N\}$), which has all features observed. The data provider only provides the values of observed variables though. In addition, it specifies a set of partitions: $\mathcal{P} = \{\mathcal{D}_x^1, \mathcal{D}_x^2, ..., \mathcal{D}_x^P\}$ where $\bigcup_{p=1}^{P} \mathcal{D}_x^p = \mathcal{D}_x$ and $\mathcal{D}_x^p \bigcap \mathcal{D}_x^q = \emptyset$ for any p, q. These partitions specify the aggregated values provided on the unobserved features ($\vec{x}_u$), $\mathcal{D}_s = \{\vec{s}_1, \vec{s}_2, ..., \vec{s}_P\}$, where $\vec{s}_p$ is derived from $\mathcal{D}_x$ as $\vec{s}_p = \frac{1}{N_p} \sum_{i=1}^{N} \vec{x}_{ui} \mathbf{1}_{(\vec{x}_{ui} \in \mathcal{D}_x^p)}$ (sample mean within $\mathcal{D}_x^p$) and $N_p = |\mathcal{D}_x^p|$. Note that in general, different partitions (and hence levels of aggregation) may apply to different unobserved variables. Though our approach can be readily extended to cover such situations, in this paper we consider a common partitioning to keep the notation and exposition simple.

Suppose we want to find K clusters in the complete data, denoted by $\{\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_K\}$. To cater to the unobserved data, for now an assumption of conditional independence) is made (to be relaxed later): $p(\vec{x}_o, \vec{x}_u|\mathcal{C}_k) = p(\vec{x}_o|\mathcal{C}_k)p(\vec{x}_u|\mathcal{C}_k)$. Let $\vec{\xi}_k$ and $\vec{\theta}_k$ be the parameters for the distributions $p(\vec{x}_u|\mathcal{C}_k)$ and $p(\vec{x}_o|\mathcal{C}_k)$ respectively. If all data features are observed at the individual level, a LDA-like clustering model can be built based on the conditional independence assumption as

Figure 1: (a) Clustering models when complete data is available (left) and (b) when only aggregates $\vec{s}$ are observed instead of $\vec{u}$ (right).



in Figure 1 (left), where $\vec{\pi}$ is sampled from a Dirichlet distribution parametrized by $\vec{\alpha}$. Figure 1 (right) shows a modified clustering model that accommodates the aggregated nature of the unobserved variables. As $\vec{x}_u$ and $\vec{x}_o$ are independent given $\mathcal{C}_k$, they can be separated using different nodes. In the model, $\vec{x}_u$ is not observed; rather the derived (aggregated) features $\vec{s}$ are observed.

Even though the model of Fig. 1(b) captures the problem characteristics, it is highly inefficient and contains redundant nodes. Fortunately, the complexity of the model can be reduced by removing the unobserved nodes $\vec{x}_u$'s if N_p is large enough. Let $\vec{\eta}_k$ and $\mathbf{T}_k$ be the mean and variance of the distribution, $p(\vec{x}_u|\mathcal{C}_k)$. Using the **linearity** of mean statistics and the **Central Limit Theorem**, $\vec{s}_p$ can be approximated as being generated from a normal distribution $\mathcal{N}(\vec{\mu}_p, \mathbf{\Sigma}_p^2)$, where $\vec{\mu}_p = \sum_{k=1}^{K} \pi_{pk} \vec{\eta}_k$, $\mathbf{\Sigma}_p^2 = \sum_{k=1}^{K} \frac{\pi_{pk} \mathbf{T}_k^2}{|N_p|}$. The generation process of $\vec{s}$ doesn't involve $\vec{x}_u$'s. As $\vec{x}_u$ doesn't contribute to the likelihood of the model, $\vec{x}_u$ can actually be removed, resulting in the efficient **C**lustering **U**sing features with **DI**fferent levels of **A**ggregation (CUDIA) model as shown in Figure 2. The generative process for CUDIA is as follows:

For $\vec{s}_p$ in $\mathcal{D}_s$,

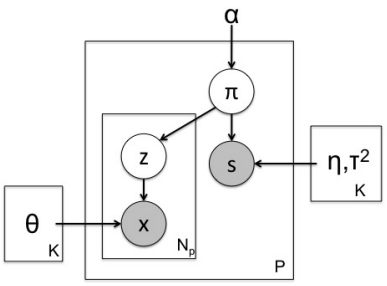
- Sample $\vec{\pi}_p \sim Dirichlet(\vec{\alpha})$.
- Sample $\vec{s}_p \sim \mathcal{N}(\vec{\mu}_p, \mathbf{\Sigma}_p^2)$,
  where $\vec{\mu}_p = \sum_{k=1}^{K} \pi_{pk} \vec{\eta}_k$ and $\mathbf{\Sigma}_p^2 = \sum_{k=1}^{K} \frac{\pi_{pk} \mathbf{T}_k^2}{N_p}$.
- For $\vec{x}_{oi}$ in $\mathcal{D}_x^p$,
  Sample $\vec{z}_i \sim Multinomial(\vec{\pi}_p)$.
  Sample $\vec{x}_{oi} \sim \prod_{k=1}^{K} p(\vec{x}_o|\vec{\theta}_k)^{z_{ik}}$.

$\vec{\pi}$ is sampled from a Dirichlet distribution parametrized by $\vec{\alpha}$, and observed sample mean statistics $\vec{s}$ is generated from a Normal distribution parametrized by a mixture of true means $\vec{\eta}$s and a covariance $\mathbf{\Sigma}^2$. $\vec{z}$'s in each partition are sampled from a Multinomial distribution parametrized by $\vec{\pi}$, which is specific to the partition, and corresponding $\vec{x}_o$s are sampled from a distribution $\prod_{k=1}^{K} p(\vec{x}_o|\vec{\theta}_k)^{z_k}$, where the suitable form of $p(\vec{x}_o|\vec{\theta}_k)$ depends on the properties of the variable $\vec{x}_o$'s. For conciseness, the remaining sections of this paper will denote $\vec{x}_o$ as $\vec{x}$.

3. INFERENCE

A generic EM algorithm [8] cannot be applied due to the coupling between $\vec{z}$ and $\vec{\pi}$ as the normalization constant of its posterior distribution is intractable. Collapsed

Figure 2: Graphical Model of CUDIA.



Gibbs sampling [13] also cannot be applied because $\vec{\pi}$ cannot be integrated out due to non-conjugacy between $\vec{s}$ and $\vec{\pi}$. However the model can be learned using either variational methods or Gibbs sampling approaches, and this paper follows the latter alternative. However, naïve Gibbs sampling approaches are computationally inefficient, thus this paper employs an approximated Gibbs sampling approach, which can be applied when the dimension of $\vec{x}$ is small. The model parameter estimation follows the MCEM algorithm [6] using this approximation technique.

3.1 E-step: Gibbs Sampling

In CUDIA, the latent variables are $\vec{\pi}$ and z. So we have:

$$p(\mathbf{X},\mathbf{S},\mathbf{\Pi},\mathbf{Z}|\vec{\eta},\vec{\theta},\vec{\alpha}) = \prod_{p=1}^{P} p(\vec{s}_p|\vec{\pi}_p,\vec{\eta})p(\vec{\pi}_p|\vec{\alpha})\prod_{i=1}^{N_p}\prod_{k=1}^{K} p(\vec{x}_i|\vec{\theta}_k)^{z_{ik}}p(\vec{z}_i|\vec{\pi}_p).$$

For each partition p, the Gibbs sampling is performed as follows:

$$\vec{\pi}_p^{(j+1)} \sim p(\vec{\pi}|\vec{z}_1^{(j)},\vec{z}_2^{(j)},...,\vec{z}_{N_p}^{(j)},\vec{s}_p,\vec{\eta},\vec{\alpha}) \quad (1)$$

$$\vec{z}_i^{(j+1)} \sim p(\vec{z}|\vec{\pi}_p^{(j+1)},\vec{x}_i,\vec{\theta}). \quad (2)$$

However, sampling $\vec{\pi}$ is problematic as Eq. (1) is not a trivial distribution. Instead of sampling directly from Eq. (1), Metropolis-Hastings (MH) algorithm can be used with a proposal density $Dirichlet(\vec{\alpha})$:

$\vec{\pi}_p^{(new)} \sim Dir(\vec{\alpha})$ and $\zeta \sim Uniform(0,1)$.

$\vec{\pi}_p^{(j+1)} \leftarrow \vec{\pi}_p^{(new)}$ if $\zeta < g(\vec{\pi}_p^{(new)},\vec{\pi}_p^{(j)})\prod_k^K(\frac{\pi_{pk}^{(new)}}{\pi_{pk}^{(j)}})^{n(z_{\cdot k}^{(j)})}$,

where $g(\vec{\pi}_p^{(new)},\vec{\pi}_p^{(j)}) = \frac{p(\vec{s}_p|\vec{\pi}_p^{(new)},\vec{\eta})p(\vec{\pi}_p^{(new)}|\vec{\alpha})^2}{p(\vec{s}_p|\vec{\pi}_p^{(j)},\vec{\eta})p(\vec{\pi}_p^{(j)}|\vec{\alpha})^2}$ and $n(z_{\cdot k}^{(j)})$ is the count of $z_{\cdot k}^{(j)} = 1$.

Even though this MH algorithm inside the Gibbs sampling becomes inefficient when dealing with large datasets, the sampling step of $\vec{z}$'s can be removed assuming a large enough data size of N_p and a small dimension of $\vec{x}$.

The overall idea of this approximation is as follows: If $\vec{x}$ is generated from an exponential family distribution, $p(z_k|\vec{x},\pi)$ is continuous with respect to $\vec{x}$, so that $p(\vec{z}|\vec{x},\vec{\pi}) \approx p(\vec{z}|\vec{x}+d\vec{x},\vec{\pi})$. Consider a ball of radius $r > 0$ centered at $\vec{x}^c$, $B_r(\vec{x}^c)$, such that $p(\vec{z}|\vec{x}^c,\vec{\pi}) \approx p(\vec{z}|\vec{x},\vec{\pi})$, where $\vec{x}$ is in the ball. If the number of $\vec{x}$'s that are in the ball is large enough, then $n(z_{\cdot k})$ in the ball can be approximated as $n(z_{\cdot k}) \approx |B_r(\vec{x}^c)|E[z_k|\pi_p,\vec{x}^c] \approx \sum_{\vec{x}\in B_r(\vec{x}^c)} E[z_k|\pi_p,\vec{x}]$. This idea can be effectively applied when N_p is large and the dimension of $\vec{x}$ is small, even better when $\vec{x}$ is a discrete variable. Assuming partitional balls over $\mathcal{D}_x^p$, $n(z_{\cdot k})$ in the partition p can be approximated as $\sum_{i=1}^{N_p} E[z_k|\pi_p,\vec{x}_i]$. Letting the number of Gibbs samples be N_{Gibbs}, the algorithm works as follows:

For $j = 1$ to N_{Gibbs},

- Sample $\pi_p^{(j+1)}$ using MH algorithm, where $n(z_{\cdot k}^{(j)}) \leftarrow \sum_{i=1}^{N_p} E[z_k|\pi_p^{(j)},\vec{x}_i]$
- Set $E[z_k|\pi_p^{(j+1)},\vec{x}_i] = \frac{p(\vec{x}_i|\vec{\theta}_k)\pi_{pk}^{(j+1)}}{\sum_{k=1}^K p(\vec{x}_i|\vec{\theta}_k)\pi_{pk}^{(j+1)}}$.

$E[z_k|\vec{x}] \propto \sum_{j=1}^{N_{Gibbs}} E[z_k^{(j)}|\pi_p^{(j)},\vec{x}]$.

The last line of the algorithm is derived by using the Partition Theorem of conditional expectation [10]. As a result, the actual sampling process occurs only in the MH sampling. In this paper, we used a burning period of 10 samples, and $N_{Gibbs} \approx 50$ to 100 [2]. Experimental results show that with this small number of samples, the algorithm converges with reasonable speed.

3.2 M-step: Parameter Estimation

Model parameters are $\vec{\alpha}$, $\vec{\theta}$ and $\vec{\eta}$. Maximization on $\vec{\alpha}$ and $\vec{\theta}$ can be easily performed and won't be discussed in this paper. $\vec{\eta}^*$ and $\mathbf{T}^*$ can be obtained by alternating the maximization steps on $\vec{\eta}$ and $\mathbf{T}$ respectively. However, if we assume $\mathbf{T}_k = \delta\mathbf{I}$ for any k, the maximization step on $\vec{\eta}$ can be simplified. To simplify the notation, the following matrices are defined [18] :

$$\mathbf{S}_i = [s_{1i}, s_{2i}, ..., s_{Pi}]^T \quad (3)$$

$$\hat{\mathbf{\Pi}} = [\hat{\vec{\pi}}_1, \hat{\vec{\pi}}_2, ..., \hat{\vec{\pi}}_P]^T, \text{ where } \hat{\vec{\pi}}_p = \frac{\sum_{i=1}^{N_{Gibbs}} \vec{\pi}_p^{(i)}}{N_{Gibbs}} \quad (4)$$

$$\mathbf{W} = diag(N_1, N_2, ..., N_P) \quad (5)$$

$$\mathbf{H} = [\vec{\eta}_1, \vec{\eta}_2, ..., \vec{\eta}_K]^T \quad (6)$$

As $\vec{s}$ is normally distributed in CUDIA, the solution of 'weighted linear regression' can be applied:

$$\mathbf{H}_{\cdot i}^* = (\hat{\mathbf{\Pi}}^T\mathbf{W}\hat{\mathbf{\Pi}})^{-1}\hat{\mathbf{\Pi}}^T\mathbf{W}\mathbf{S}_i. \quad (7)$$

Note that $rank(\hat{\mathbf{\Pi}}^T\mathbf{W}\hat{\mathbf{\Pi}}) = rank(\hat{\mathbf{\Pi}}) = K$ w.p. 1 if $P > K$. However, mean values ($\hat{\mathbf{\Pi}}$) are susceptible to outliers from the Gibbs sampling. To ensure the invertibility, regularization techniques can be incorporated. For example, if a Ridge penalty is used, then $\mathbf{H}$ becomes:

$$\mathbf{H}_{\cdot i}^* = (\hat{\mathbf{\Pi}}^T\mathbf{W}\hat{\mathbf{\Pi}} + \lambda\mathbf{I})^{-1}\hat{\mathbf{\Pi}}^T\mathbf{W}\mathbf{S}_i. \quad (8)$$

4. DETERMINISTIC HARD CLUSTERING

The CUDIA model provides an intuitive deterministic hard clustering algorithm. From the log-likelihood of CUDIA, the objective function becomes:

$$\min_{\mathbf{Z},\vec{\mu},\vec{\eta}} \sum_p \{\sum_{k,n_p} z_{n_pk} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2\} + \beta \parallel \vec{s}_p - \sum_k \frac{\sum_{n_p} z_{n_pk}}{N_p}\vec{\eta}_k \parallel^2 \quad (9)$$

$$= \min_{\mathbf{Z},\vec{\mu},\vec{\eta}} \sum_{p,k,n_p} z_{n_pk} \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 + \frac{\beta}{KN_p} \parallel \vec{s}_p - \sum_k \hat{\pi}_{pk}\vec{\eta}_k \parallel^2 \quad (10)$$

Table 1: Dataset Description. Target is not included when performing the imputation. The top 5 biggest population states are selected to maintain large enough N_p. (Case 1)

	Hospital-level
1	Hospital beds(**Target**)
2	Home health agency visits per decedent
3	Percent of deaths occurring in hospital

	State-level
1	Medical discharge rate
2	Surgical discharge rate

where $\hat{\pi}_{pk} = \frac{\sum_{n_p} z_{n_p k}}{N_p}$ and β is a parameter that determines weights to mean statistics. Local minima of this objective function can be found by alternating minimization steps between $\mathbf{Z}$ and $(\vec{\mu}, \vec{\eta})$:

- **Assignment Step**

  $z_{n_p k^*} \leftarrow 1$,
  if $k^* = \arg\min_k \parallel \vec{x}_{n_p} - \vec{\mu}_k \parallel^2 -2(\vec{s}_p - \mathbf{H}^T \hat{\vec{\pi}}_p)^T \vec{\eta}_k (\frac{\beta}{K N_p})$

  $z_{n_p k^*} \leftarrow 0$, otherwise.

- **Update Step**

  $\vec{\mu}_k \leftarrow \sum_n (z_{nk} \vec{x}_n)/N_k$, $\vec{\pi}_p \leftarrow \sum_{n_p} \vec{z}_{n_p}/N_p$

  $\mathbf{H}_{\cdot i} \leftarrow (\hat{\mathbf{\Pi}}^T \mathbf{W} \hat{\mathbf{\Pi}} + \lambda \mathbf{I})^{-1} \hat{\mathbf{\Pi}}^T \mathbf{W} \mathbf{S}_i$

5. EXPERIMENTAL RESULTS

In this section, several experimental results using the CUDIA model are provided using data from the Dartmouth Health Atlas, CDC and the Census Bureau. Depending on the nature of the predictor and the data source, averaged values are provided at hospital, county, HRR/HSA or state levels. Thus "individual" will refer to either a single hospital or a single county as these are at the finest granularity level in the corresponding studies. The CUDIA model is used to impute the aggregated features at the individual-level, and its results are compared to predictive modeling using only higher level data.

5.1 Dartmouth Health Atlas: Case 1

The Dartmouth Health Atlas dataset [16] is composed of several tables with different levels of aggregation. For example, the number of beds in a hospital can be accessed at the hospital-level, whereas the medical/surgical discharge rates can only be obtained at State/HRR/HSA levels. Table 1 describes the subset of the Dartmouth data used in this experiment. Only data from the 5 most populous states (CA, FL, IL, NY, TX) was used so as to have a higher value of number of hospitals per state. For this subset, the "complete data" would have consisted of five variables at the hospital-level, of which two are actually available only at the state level. The CUDIA model can be used to impute the unobserved features ($\vec{x}_u$). The imputation can be performed as follows:

$$\hat{\vec{x}}_u \leftarrow \sum_{k=1}^{K} E[z_k | \vec{x}_o] \vec{\eta}_k \tag{11}$$

Table 2: Regression Results on Dartmouth datasets. R^2s over 5-fold cv are listed. As $K < P$, $K > 5$ is not an option.

Dataset	Case 1	Case 2	Case 3
No Imputation	0.548	0.547	0.667
	(±0.056)	(±0.030)	(±0.026)
State-level Imputation	0.559	0.576	0.671
	(±0.061)	(±0.037)	(±0.029)
CUDIA Imp. ($K = 2$)	0.557	0.539	0.659
	(±0.052)	(±0.032)	(±0.052)
CUDIA Imp. ($K = 3$)	0.552	0.545	0.680
	(±0.056)	(±0.032)	(±0.023)
CUDIA Imp. ($K = 4$)	0.563	0.593	**0.686**
	(±0.057)	(±0.029)	(±0.027)
CUDIA Imp. ($K = 5$)	**0.577**	**0.596**	0.684
	(±0.056)	(±0.030)	(±0.027)

Table 3: Coefficients of Linear regression when $K = 5$. All features are standardized. (Case 1)

Independent Variable	Coefficient
Home health agency visits per decedent	0.371
Percent of deaths occurring in hospital	0.928
Medical discharge rate	**-0.180**
Surgical discharge rate	**0.376**

Linear regression is used to perform the task on three kinds of datasets: 1) 'hospital-level' dataset alone, 2) imputed complete dataset using 'state-level' summaries and 3) imputed dataset using the CUDIA model. 5-fold CV is performed and Table 2 shows the results. $K = 5$ gives the best R^2 value among all the alternatives. Table 3 shows the coefficients of Linear regression when $K = 5$. The imputed medical discharge rate is negatively correlated with the number of beds in a hospital.

5.2 Dartmouth + External Source: Case 2

State-level summaries of health-related indicators can be obtained from various external sources. For example, the Center for Disease Control and Prevention (CDC) publishes annual state-level health statistics, that covers aging, cancer, diabetes, etc. In this experiment, the Dartmouth dataset is used with an external dataset from StateMaster.com, which provides multiple state-level statistics for free. The hospital-level Dartmouth dataset from the previous experiment is used as is. The state-level dataset is replaced with the external dataset, which has state-level 1) healthcare spending, 2) hospital admissions and 3) adult physical disabilities information. All these features are not available in the Dartmouth data. As in the previous experiment, three datasets are formed. Table 2 shows the R^2 results using 5-fold CV. Imputation using the CUDIA model leads to a 9% increase in R^2 value compared to the base model without imputation. Table 4 shows the coefficients of Linear regression when $K = 5$. The imputed healthcare spending exhibits the strongest correlation with hospital spending, as one may expect.

Table 4: Coefficients of Linear regression using the external source when $K = 5$. All features are standardized. (Case 2)

Independent Variable	Coefficient
Home health agency visits per decedent	0.296
Percent of death occurring in hospital	0.290
Healthcare spending	**0.382**
Admissions	**0.211**
Adult physical disabilities	**-0.102**

Table 5: Coefficients of Linear regression using the external source when $K = 4$. All features are standardized. (Case 3)

Independent Variable	Coefficient
Medicare Part-B	0.548
Income per capita	**0.018**
Healthcare spending	**-0.154**
Education level (Bachelor or higher)	**-0.329**

5.3 Dartmouth + External Source: Case 3

In this experiment, Medicare part-A reimbursement at HSA-level is predicted based on Medicare part-B reimbursement and an additional external information. In the Dartmouth dataset, 'Selected Medicare Reimbursement' table contains the columns of Medicare reimbursement part-A and part-B at HSA-level. Although Medicare part-A is closely related to part-B, additional features, such as income or education levels, can be incorporated not only improving the performance of the regression but also providing richer interpretations. The external state-level features used in this experiment are 1) income per capita, 2) total healthcare spending and 3) education level (ratio of bachelors or higher). The experiment is performed using three datasets, which are prepared as in the previous experiments. Table 2 shows the results. Table 5 exhibits the coefficients of Linear regression when $K = 4$. The imputed 'education level' and Medicare Part-A are negatively correlated.

5.4 CDC Diabetes Dataset

The Center for Disease Control and Prevention (CDC) [7] provides county-level estimates of 1) obesity, 2) diabetes and 3) physical inactivity. In this experiment, we predict the county-level obesity rate using the other features in the CDC dataset and additional state-level features. The state-level features used in this experiment are the same as in the previous experiment (Dartmouth Case 3). The top 5 biggest states are used, as some smaller states have very few counties. Table 7 shows the R^2 results. The imputed dataset using the CUDIA model gives the best result. The state-level imputed dataset yields a poorer result than the dataset with no imputation. This indicates that the uncertainty in the state-level imputation of the added variables over-rode any extra benefits that these variables could have provided. Table 6 depicts the coefficients when $K = 5$. While the imputed 'income per capita' at county-level shows a negative correlation, both imputed 'healthcare spending' and 'education level' are positively correlated with the target (obesity rate at county-level).

Table 6: Coefficients of Linear regression on CDC diabetes dataset when $K = 5$. All features are standardized. 'Obesity rate' is set as a target.

Independent Variable	Coefficient
Diabetes	0.019
Physical inactivity	1.29
Income per capita	**-0.224**
Healthcare spending	**0.227**
Education level (Bachelor or higher)	**0.388**

Table 7: Regression Results on CDC Diabetes and Census Bureau Dataset. R^2s over 5-fold cv are listed.

Dataset	CDC Diabetes	Census Bureau
No Imputation	0.401 (±0.034)	0.508 (±0.056)
State-level Imputation	0.395 (±0.034)	0.506 (±0.056)
CUDIA Imp. ($K = 2$)	0.347 (±0.215)	0.505 (±0.20)
CUDIA Imp. ($K = 3$)	0.296 (±0.033)	0.499 (±0.055)
CUDIA Imp. ($K = 4$)	0.419 (±0.035)	0.514 (±0.058)
CUDIA Imp. ($K = 5$)	**0.421** (±0.032)	**0.520** (±0.060)

5.5 Census Bureau Health Insurance Dataset

The U.S. Census Bureau [17] provides county-level estimates of insured population ratio by income levels. Income levels are divided into three overlapping groups: 1) all income levels, 2) at or below 200% of poverty threshold and 3) at or below 250% of poverty threshold. Suppose we want to see which other factors affect propensity of poor people to buy healthcare insurance at the county level. The state-level dataset in the previous experiment is used to determine if other factors play a role. Table 7 shows the regression results using the CUDIA model and the coefficients when $K = 5$ are described in Table 8. 'Income per capita' and 'education level' are negatively correlated with the target (percent insured for the below 200% of poverty group). This result indicates that the imputed county-level summaries for both income per capita and education level implicitly inform us of the sizes of poverty group at county-level. Moreover, the imputed healthcare spending at county-level exhibits a positive relationship. Thus these imputed features provide a richer interpretation of the predictive model while simultaneously improving the prediction accuracy.

6. CONCLUDING REMARKS

In this paper, aggregated statistics over certain partitions are utilized to identify clusters and impute features that are observed only as more aggregated values. The imputed features are further used in Regression modeling, leading to im-

Table 8: Coefficients of Linear regression on Census Bureau dataset when $K = 5$. All features are standardized. 'Percent insured for the below 200% of poverty' is set as a target.

Independent Variable	Coefficient
Percent insured for all income levels	2.10
Income per capita	**-0.436**
Healthcare spending	**1.227**
Education level (Bachelor or higher)	**-2.524**

proved R^2 values. The experiments provided in this paper are illustrative of the generality of the propsed framework and its applicability to several healthcare related datasets in which individual records are often not available, and different information sources reflect different types and levels of aggregation. Empirical studies on larger and richer datasets are forthcoming.

CUDIA is quite scalable, and in particular, the deterministic hard clustering version of the CUDIA model can be readily applied to massive datasets. Furthermore, the square loss function on $\vec{x}_o$ can be generalized to Bregman divergence, or equivalently, one can cater to any noise function from the exponential family of probability distributions [4]. One restriction of the current model is that the number of clusters (K) cannot be more than the number of partitions specified by the data provider(P). This is why we had to stop at K=5 for several of the results even though the R^2 values were improving with with increasing K. Adding more partitions, e.g., incorporating data from more than 5 states, should reflect in further improvements in the results.

Acknowledgments

This research was supported by NSF IIS-1016614.

7. REFERENCES

[1] Texas Inpatient Public Use Data File. *http://www.dshs.state.tx.us/thcic/hospitals /HospitalData.shtm.*

[2] D. Agarwal and B. Chen. Regression-based latent factor models. In *KDD '09*, pages 19–28, 2009.

[3] R. Agrawal and R. Srikant. Privacy-preserving data mining. In *ACM SIGMOD*, pages 439–450, 2000.

[4] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Jl. Machine Learning Research (JMLR)*, 6:1705–1749, October 2005.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pages 993–1022, 2003.

[6] J. G. Booth and J. P. Hovert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B*, 61:265–285, 1999.

[7] Centers for Disease Control and Prevension (CDC). *http://apps.nccd.cdc.gov/DDTSTRS/default.aspx.*

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[9] C. Dwork. Differential privacy. In *ICALP*, volume 4052, pages 1–12, 2006.

[10] G. Grimmett and D. Stirzaker. *Probability and Random Processes*, chapter 3.7, page 67. Oxford, third edition, 2001.

[11] C. Jackson, N. Best, and S. Richardson. Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors. *Journal of Royal Statistical Society: Series A*, 171:159–178, 2008.

[12] C. Jackson, N. Best, and S. Richardson. Bayesian graphical models for regression on multiple data sets with different variables. *Biostatistics*, 10(2):335–351, 2009.

[13] J. S. Liu. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966, 1994.

[14] L. Sweeney. Information explosion. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, 2001.

[15] H. T. Tavani. Information privacy, data mining, and the internet. In *Ethics and Information Technology*, volume 1, pages 137–145, 1999.

[16] The Dartmouth Atlas of Health Care. *http://www.dartmouthatlas.org/.*

[17] U.S. Census Bureau. *http://www.census.gov/did/www/sahie/data /2007/dataset.html.*

[18] G. C. G. Wei and M. A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.