Predicting from Aggregated Data

By
Maaz Babekir Ali Jedh

Senior Thesis in Computer Engineering
University of Illinois at Urbana-Champaign

Advisor: Prof. Sanmi Koyejo

December 2020
Abstract

Aggregated data, which refers to a collection of data summarized from multiple sources, is a technique commonly used in different fields of research including healthcare, web application, and sensor network. Aggregated data is often employed to handle issues such as privacy, scalability, and reliability. However, accurately predicting individual outcomes from grouped datasets can be very difficult. In this thesis, we designed a new learning method, a Mixture of Expert (MoE) model, focused on individual-level prediction when training variables are aggregated. We utilized the MoE model, trained and validated using the eICU Collaborative Research patient datasets, to conduct a series of studies. Our results showed that applying grouping functions to the classification of aggregated data across demographic and behavior metrics could remain effective. This technique was verified by comparing two separately trained MoE models that were evaluated on the same datasets. Finally, we estimated non-aggregated datasets from spatio-temporal aggregated records by expressing the problem into the frequency domain, and trained an autoregressive model for predicting future stock prices. This process can be repeated, offering a potential solution to the issue of learning from aggregated data.

Keywords: Mixture of Experts, Data Aggregation
1 Introduction:

Machine learning methods have enhanced the state-of-the-art technologies used to reshape social patterns, booting computational abilities that pertinent many facets in the modern world. Increasingly, these innovations make use of a subset of machine learning known as deep learning. The computational challenges of deep learning allow scientists to build many tools that exploit hardware features, such as multi-central processing units (CPUs) and graphics processing units (GPUs), that shorten the training and inference time. Building and training deep learning algorithms alongside their mathematical properties has empowered multi-billion dollar industries. Moreover, combining such advancements with the high performance of deep neural networks has helped overcome numerous big data challenges (Shi et al., 2016). Despite the increasing emergence of applications that train on individual datasets across different domains, accessing raw data is restricted due to privacy reasons and training datasets are aggregated (Goodfellow et al., 2016; Shen et al., 2017; Bhowmik et al, 2019). This presents an opportunity for researchers to analyze aggregated data and use machine learning tools to extract individual-level prediction.

Learning from aggregated data requires a novel technique that exploits significant deep learning algorithms and data aggregation strategies. There are situations that impose a need for applying other mathematical methods such as integration by parts for spatial-temporal datasets to compress raw datasets, which must be utilized while investigating questions within advertising, healthcare, education, and real estate businesses just to list a few examples.

One data aggregation use is for privacy preservation; data aggregation has been used in smart grid power management as a strategy towards protecting regional residents’ power consumption records. Grouping daily electricity use from specific regions hides daily home power consumption records, which is subject to regulation and ethics guidelines, restricting companies
from publicly revealing residential information and preventing hackers from tracking consumer behavior such as living habits and lifestyles (Shen et al., 2017). A similar strategy is also applied in healthcare (e.g. patient records), education (e.g. SAT exams score), and social media (e.g. user comment history).

Another data aggregation use is for sharing and storing large volumes of data; transporting and storing big data over large distances is tremendously complex and hard to maintain. In this case, data aggregation plays a major role by grouping data prior to transferring, and applying big data reduction methods to precompute storage size; this achieves high performance querying time for large datasets (Shen et al., 2017; Rehman et al., 2016). This can be seen in figures released by the Bureau of Labor Statistics and Amazon's storage management and data production (US Department of Labour; Engdahl et al., 2008).

1.1 Mixture of Expert Model:

In 1991, Jordan and Jacobs introduced the mixture of experts' approach, an architecture that is based on the divide-and-conquer principle and multiple neural network experts (learners). In this approach, data is divided into subsets that are trained separately, forming subgroups composed of multiple learners. In addition to a gating network that supervises which experts are used for each training set, resulting in an increase in model capacity without a proportional increase in computation time. Best performing learners are rewarded with a strong feedback signal and specialized for inputs with similar patterns (Yukse et al., 2012; Masoudnia et al., 2014; Krzysztof et al., 2017; Jacobs et al., 1991). In this thesis, we will be dividing aggregated data into subgroups, where each learner is trained on a subgroup. This a natural mode of human thought, where big problems are broken into smaller pieces and combined together for a solution. The MoE approach
has been applied over the past 20 years to solve complex machine learning problems from face recognition and surveillance to healthcare and finance. Learning from aggregated data using MoE algorithms could help minimize ecological fallacies, which occur when making a faulty individual conclusion based on the analysis of collective information (Ebrahimpour et al., 2007).

Figure 1: A Mixture of Experts (MoE) layer, in which, the sparse gating function selects multiple experts to perform computations. Their outputs are modulated by the outputs of the gating network (Krzysztof et al., 2017).
1.2 Spatio-temporal Data:

In the past five decades, the use of spatio-temporal applications has increased rapidly, requiring the development of a fast robust algorithm, and optimal solutions to learn from these practical applications. Individual level prediction tasks from time aggregated spatio-temporal datasets are affected by several complex patterns, such as cities, weather, and holidays (Xie et al., 2014). Raw datasets are averaged across time with at least one spatial and one temporal property (Ferreira et al., 2020). A spatio-temporal dataset represents a phenomenon at a certain time and location, in which it introduces a new study technique investing persistence and unusual patterns over time. An example of such datasets would be information reported by the Center for Disease Control and Prevention regarding Coronavirus Disease 2019 (COVID-19) patients' admitted to a hospital each day between March and August 2020 (CDC).
2 Related Work:

Learning from aggregated data is a semi-supervision problem that has not been studied extensively in the past; there have been very few successful attempts to utilize training with grouped data and individual analyses. Mixed individual and aggregated records in a clustered setting, generatively processed using the Bayesian direct graphical model, capturing the properties of aggregate level data using the Central Limit theorem to estimate raw data prediction through the application of an approximated Gibbs sampling method (Park et al., 2012). The authors later introduce the LUDIA method, an algorithm that estimates original individual level values from aggregated data by utilizing aggregation constraints and auxiliary information. The LUDIA algorithm was able to reconstruct individual data from grouped records across county, hospital, and zip code levels. Yousefi et al. (2019) proposed a multi-task learning model based on Gaussian processes, that approximates raw data sets that have been aggregated at different input scales. Developing a model that looks at each task as a linear combination of scaled latent process realization allows the individual task to be assigned a likelihood model providing a variational, which can be optimized for learning. Bhowmik et al. (2019) introduced a framework that uses aggregated data for building individual-level predictive models by applying generalized linear modeling.

On the other hand, the mixture-of-experts approach has been subject to enormous development and research over the last two decades. New types of MoE architecture have been developed. For instance, the Gaussian Process where the MoE model’s experts were replaced with regression models, which is useful for input dependent applications (Volker et al., 2001). Bangpeng et al. (2009) introduced a different expert configuration to model connections in the Hidden Conditional Random Field framework. Here the classifier of one region of interest makes
predictions based on classifier predictions from connected regions of interests. Similarly, Krzysztof et al. (2017) brought forward a Sparsely Gated Mixture of Experts Layer model that has outperformed traditional deep learning networks, achieving higher over 1,000x improvement in model capacity on modern GPU clusters, which consist of updating thousands of feed-forward sub-networks. Coming up with such architecture has allowed a lower computational cost when training using an MoE model while still yielding high prediction accuracy.

In this paper, a novel algorithm known as Frequency Domain Predictive Modelling with Aggregated Data, designed by Bhowmik et al. (2017), will be used. This method uses given individual aggregated samples at a certain period, which is equivalent to a convolution operation with a square and sampling. The method then constructs datasets into the frequency which captures global aggregation domain in local time, and divides the frequency by a sinc function to retrieve non-aggregated data estimates. By exploiting the duality properties of Fourier analysis for spatio-temporal aggregated datasets, that are collected in a non-uniform method across targets and features, they were able to estimate a non-aggregated dataset. This proposed method is described in figure[2]. Further, the Fourier Transform deconvolution for electron paramagnetic resonance imaging (EPRI) pressed by Deng et al. (2003) will also be applied to determine the cutoff frequency by calculating the piecewise variance of the division result of the Fourier amplitude spectra.
**Algorithm 1** Fourier-domain Estimation from Aggregated Data

1: **Input:** $\bar{x}, \bar{y}, \omega_0, D, T_0$

2: Sample $D$ frequencies uniformly in $(-\omega_0, \omega_0)$ to get
   $\Omega = \{\omega_1, \omega_2, \cdots, \omega_D : \omega_i \in (-\omega_0, \omega_0)\}$

3: **for** each $\omega \in \Omega$, and $i \in \{1, 2, \cdots, d\}$ **do**

4: compute the $T_0$-limited finite Fourier transforms
   $\bar{X}_{i,T_0}(\omega) = \mathcal{F}_{T_0}\bar{x}_i(\omega)$,
   $\bar{Y}_{T_0}(\omega) = \mathcal{F}_{T_0}\bar{y}(\omega)$

5: reconstruct non-aggregated Fourier Transforms
   $\hat{\bar{X}}_{i,T_0}(\omega) = \frac{\bar{X}_{i,T_0}(\omega)}{U_{T_1}(\omega)}$,
   $\hat{\bar{Y}}_{T_0}(\omega) = \frac{\bar{Y}_{T_0}(\omega)}{U_T(\omega)}$

6: **end for**

7: Estimate the parameter as
   $\hat{\beta} = \arg \min_{\beta} \frac{1}{|\Omega|} \sum_{\omega \in \Omega} ||\hat{\bar{X}}_{T_0}(\omega)^\top \beta - \hat{\bar{Y}}_{T_0}(\omega)||^2$

8: **return** $\hat{\beta}$

---

Figure 2: Fourier-domain Estimation from Aggregated Data Algorithm (Bhowmik et al., 2017);
3 Methods:

This section describes the data curation methodology followed to prepare data sets for training and evaluation. Patients’ data sets were acquired from the eICU Collaborative Research database (Pollard et al., 2018), which was then aggregated for this analysis - we provide all open source information and aggregation functions described for reproducibility purposes. Then, we describe the mixture of expert network models used to classify allergies using raw and aggregated datasets. Finally, we use an aggregated version of the individual (open & close stock price) Apple Stock price obtained from Yahoo! Finance for estimating daily non-aggregated stock price and forecasting future stock price.

3.1 Data Curation:

The eICU Collaborative Research database is a multi-center database that was developed from healthcare data for over 200,000 intensive care unit (ICU) admissions across the United States between 2014 and 2015. The database includes over 250,000 allergy diagnoses, where some subjects are diagnosed with multiple allergies. For this experiment, we chose the top five allergies for classification, whose distribution across 13,474 patients is shown in Table 1. Then we computed aggregates with the same datasets, this allows us to have two different datasets used for training and evaluating our results by comparing both MoE models that are trained with aggregate and individual datasets.
We perform a basic aggregation technique across hospitals to preserve patients' privacy, in which datasets were impossible to trace back to individual patients. To de-identify and protect anecdotal datasets we followed a method suggested by the US Department of Health & Human Services, and we applied count, maximum, sum, and average aggregation functions. Our classification depends on seven features, described in Table 2. Patients are aggregated by identical allergy types and hospital-id, where age and height are averaged and normalized. Then, each group ethnicity and gender are labeled using One-hot averaging method. As a result, 13,474 patients' were reduced to 143 groups.

3.2 Mixture of Expert (MoE) Model:

Deep learning has provided promising results for numerous applications in natural language processing, object detection, and especially in classification. For further improvement, exploiting existing deep learning architecture lessens the sophistication of surrounding machine learning methods (Wang et al., 2015). This experiment follows the premise of this paradigm by applying the MoE model with aggregated data, which was designed to improve model capacity, training time, and/or model quality (Krzysztof et al., 2017). In this experiment, our MoE consists

<table>
<thead>
<tr>
<th>Allergy Types</th>
<th>Number of Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cerner</td>
<td>5,741</td>
</tr>
<tr>
<td>Codeine Phosphate</td>
<td>2,235</td>
</tr>
<tr>
<td>Benzathine Benzylpenicillin</td>
<td>2,133</td>
</tr>
<tr>
<td>Morphine Sulfate</td>
<td>2,068</td>
</tr>
<tr>
<td>Iodine</td>
<td>1,297</td>
</tr>
</tbody>
</table>
of multiple neural networks with a single hidden layer and controlled through a gating network, shown in Figure 1, where the gating network determines a sparse combination used by each expert for each training sample.

The main distinction between existing work in this domain, of learning from aggregate, is that our approach uses multiple neural networks and a gating network to estimate individual level prediction. Learning parameters consists of two tasks that learn from individual experts and a gating network. While each expert competes on computing inputs, a getting network mediates their result. The final prediction of the combined network is a weighted average (with expert weights \( w_1, w_2, w_3, \ldots, w_n \)) of the outputs \( (y_1, y_2, y_3, \ldots, y_n) \) of each expert network. In our approach, we observe that classification with our model performs at its best and ensures patient privacy when individual data sets are normalized.

To observe our learning efficiency, we built two MoE models, Raw model and Aggregated model, each with a different size batch, 500 and 10 respectively. An 80-20 train-validation split strategy was used in the Raw model, and the same validation and test data sets were used to compare the Aggregated model’s performance to Raw model’s performance. Our MoE models were trained using PyTorch. In section 4, we provide statistical measures from the average, mean, and variance results of our training.

### 3.3 Non-Aggregated Data Estimate:

As aspiration in spatio-temporal research, we apply Fourier-domain Estimation from Aggregated Data Algorithm to a daily aggregated stock prices. We obtained Apple's stock prices from Yahoo! Finance which includes high, low, open, and close value for each day, since Apple's
first initial public offering IPO. For this experiment, we used the mean aggregation function; the average between close and open stock price, as result, our data are grouped across one feature. Exploiting the frequency domain helps extract high fidelity estimates of individual global properties in the time domain Figure 3. During training, we randomly select a frequency sampled value within our training period. We presented a new approach to learn from aggregated spatio-temporal data using estimated non-aggregated datasets and applying an autoregressive model for prediction.

1. $\hat{z}_i[l] = \frac{1}{T_i} \int_{(t-1)T_i/2}^{lT_i/2} z_i(\tau) \, d\tau$

2. $\hat{Z}_T(\omega) = \mathcal{F}[z](\omega) = \int_{-T}^{T} \hat{z}(t) e^{-i\omega t} \, dt$

3. Sample frequencies uniformly between $(-\omega_0, \omega_0)$

   \[
   \{\omega_1, \omega_2, \omega_3, \omega_4, \ldots \colon \omega_i \in (-\omega_0, \omega_0)\}
   \]

   \[
   U_{T_i}(\omega) = \frac{\sin(\omega T_i/2)}{\omega T_i/2}
   \]

4. Estimate non-aggregated Fourier transform:

   \[
   Z_{i,T_0}(\omega) = \frac{\hat{Z}_{i,T_0}(\omega)}{U_{T_i}(\omega)}
   \]

Figure 3: Non-aggregated data estimation procedure optimized from the novel algorithm described in Figure 2; to handle one input feature aggregation.
4 Results:
As mentioned before, we evaluate our classification on two MoE models; the first model trained with raw data and the second model trained with aggregated data. The accuracy classification for each model is detailed below.

4.1 Individual datasets:
We perform raw data prediction using 13,474 datasets with age, height, gender, and ethnicity features. We split our data sets into 12 batches with a size of 1000 patients per batch, then we iterated through to see which iteration leads to a higher prediction. Having multiple iterations shows the resilience and robustness of the MoE model because we start with all zero weight and gradually use previous weights for training.

Prediction Accuracy = 60.38%

<table>
<thead>
<tr>
<th></th>
<th>Cerner</th>
<th>Codeine Phosphate</th>
<th>Benzathine G</th>
<th>Morphine Sulfate</th>
<th>Iodine</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1-Score</td>
<td>0.97</td>
<td>0.38</td>
<td>0.26</td>
<td>0.36</td>
<td>0.22</td>
</tr>
<tr>
<td>Precision</td>
<td>0.95</td>
<td>0.32</td>
<td>0.31</td>
<td>0.34</td>
<td>0.46</td>
</tr>
<tr>
<td>Recall</td>
<td>0.99</td>
<td>0.47</td>
<td>0.24</td>
<td>0.36</td>
<td>0.17</td>
</tr>
</tbody>
</table>

In this section, we exploit one main reason for this experiment is to use conditional computation for prediction along with improving model capacity. Conditional computation plays a major role in the MoE model, where part of the network is active on a per-input basis. The
performance of this model shows that part of the network that's active in training and validation is an expert on certain features.

### 4.2 Aggregate datasets:

The datasets used in this training are only containing 143 patient records, aggregated from 13,474 patient records used in training the previous exercise. This training also performed multiple times to further investigate how vulnerable our result to the ecological fallacy. The results in Table 4 shows the accuracy rate predictions from an MoE model on the same datasets.

**Prediction Accuracy** = 47.55%

<table>
<thead>
<tr>
<th></th>
<th>Cerner</th>
<th>Codeine Phosphate</th>
<th>Benzathine G Benzylpenicillin</th>
<th>Morphine Sulfate</th>
<th>Iodine</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>F1-Score</strong></td>
<td>0.79</td>
<td>0.36</td>
<td>0.20</td>
<td>0.25</td>
<td>0.15</td>
</tr>
<tr>
<td><strong>Precision</strong></td>
<td>0.89</td>
<td>0.27</td>
<td>0.23</td>
<td>0.34</td>
<td>0.13</td>
</tr>
<tr>
<td><strong>Recall</strong></td>
<td>0.71</td>
<td>0.53</td>
<td>0.18</td>
<td>0.20</td>
<td>0.18</td>
</tr>
</tbody>
</table>

We investigate our result using prediction from individual data, group-wise prediction, the best performing is the MoE model trained with raw datasets.
4.3 Apple Stock Price Raw Data Estimated:

In this section, we present raw data estimation for averaged Apple's stock price daily, Figure 3 shows both aggregated and estimated stock price. We evaluate our result by visuals and comparing between our non-aggregated and true datasets, open/close, Figure 4. However, we do generate multiple estimated raw datasets, due to the random selection of our sampled frequency(-\(\Pi/4, \Pi/4\)).

![Estimated Stock Price vs Open Stock Price](image)

Figure 4: Estimated Apple stock price sample output using simplified Frequency domain predictive modelling with aggregated data algorithm, described in Figure 3.
4.4 Autoregressive Model Result:

Here, we performed future forecasting for Apple stock prices using an ARIMA model. Which depends on past time series values, to overcome stationarity within stock price datasets, we normalize our data using a Logarithm function. We evaluate our result by visuals and comparing the ARIMA model future 5 days forecasting between our best generated non-aggregated and true datasets; open stock price, Figure 5.

Figure 5: a Sample output of an ARIMA model trained with (a) open and (b) non-aggregated estimate.
5 Discussion:

In this paper we conducted a detailed analysis of the MoE model architecture with aggregated and non-aggregated datasets estimation with frequency domain. In general, learning from grouped data cannot be guaranteed to be close to learning with individual datasets. Our technique is similar to other work that achieves optimal learning strategy in machine learning to produce a model capable of classifying allergies while maintaining patient privacy.

Another unique approach discussed in this paper is our aggregation strategy. We focused our procedure on preserving patients' privacy while maintaining global individual datasets property, as a result, a 13% accuracy gap between two MoE models was achieved. Models were separately trained with individual datasets and aggregated datasets, and separately evaluated. Secondly, estimating non-aggregates using daily grouped stock prices by leveraging Frequency domain properties and Restricted Fourier Transforms. While specifying a smaller period of integration to randomly and choose a sampling frequency. This work shows the state of the art tools in machine learning that can be used to advance learning from aggregated data, and achieve a near prediction accuracy as training from anecdotal data.

Limitations:

The shortage of labeled data is one of the key limitations to maximize the performance of our MoE models and increase our prediction accuracy. We suggest obtaining large datasets. On the other hand training with a small database illustrates that learning from aggregate data is achievable. Another challenge we faced was predicting from aggregated datasets is subject to Ecological Fallacy, indicating that predicting from aggregated datasets uniquely differs from predicting from individual datasets (Wagner et al., 1982). To correctly demonstrate our learning approach we compare our classification result with another MoE model trained with raw datasets.
We are currently searching on obtaining a large database and applying different aggregation procedures in preserving patient privacy to further enhance our prediction from aggregated data.
6 Conclusion:
The complexity of predicting from aggregated data is an important issue that needs to be investigated. In this paper, we introduced a new learning approach predicting from aggregated datasets using an existing deep learning algorithm and estimating non-aggregated datasets from spatial-temporal aggregated data using Frequency-domain property. Our findings indicate that we can achieve an accuracy rate that is near to predicting from a model trained with raw datasets with an experiment on data from healthcare. Through our investigation, we showed that deep learning algorithms perhaps be used to learn and enrich our understanding when accessing individual datasets at ground truth level are impossible. Our result suggests the importance of training with a deep learning architecture that mimics the human brain learning process in breaking aggregation dependencies, which was originally designed to enhance computational performance. Surely, our strategy for predicting from aggregate data reveals an accuracy rate that is close to training with raw datasets, and our model estimated non-aggregated spatio-temporal using frequency domain representation of spatio-temporal aggregates. Keeping in view the analysis of this paper, we look forward to discovering many novel techniques that increase our learning from aggregated data.
References:


A Bhowmik, “Learning from Aggregated Data February” 2019, doi: 2152/74150


