Asymptotically-Stable Privacy Preservation Technique for fMRI Data Privacy Protection

Naseeb Thapaliya, Lavanya Goluguri, and Shan Suthaharan

Department of Computer Science
University of North Carolina at Greensbo ro
Greensboro, NC, 27412, USA
{n_thapal,l_golugu,s_suthah}@uncg.edu

Abstract

This paper presents a computational technique that performs an asymptotic-stabilization of transition probabilities, characterized by two-state Markov chain, to protect the privacy of the functional magnetic resonance imaging (fMRI) data. The fMRI data reveal a large number of correlated brain features that can be utilized in the development of predictive models for extracting brain networks and infer privacy information of an individual. These features make fMRI data highly vulnerable to privacy attacks. To protect them for privacy enhancement, we transform them to an asymptotic state using the concepts of asymptotic stabilization with two-state Markov chain, and the compressed sensing and compressed learning techniques. The proposed predictive model is built using the asymptotically-stabilized fMRI signals, rather than the original signals, which enhances the protection of individual's privacy. The computer simulations demonstrate that the proposed predictive model provides very high prediction accuracy scores, while providing very strong privacy protection.

Introduction

The asymptotic analysis is a method of characterizing the limiting behavior of a predictive model in a wide variety of domains to improve and establish validity of the models. Mathematically, if a model is described by a simple function $f(n)$, then the asymptotic analysis is to study the properties of the function $f(n)$ at very large value of the variable n (i.e. $n \to \infty$). It is important to characterize the asymptotic stabilization of any model, since it can help us evaluate both the efficiency and effectiveness of a predictive model. Furthermore, an equilibrium condition defines the stationary conditions for the dynamics of the model and helps to determine the asymptotic stability. We adapt the concept of asymptotic stabilization to design our predictive models for protecting the privacy of fMRI data. Let's first look at why fMRI data is so crucial that it merits high level of protection from external attacks. The current advances in artificial intelligence allows the development of efficient computational techniques [1] that are capable of discovering complex structures and functionalities of brain networks from fMRI signal and machine learning models.

The brain network is unique to an individual; hence, the extraction of structural (revelation of region of interests) and functional brain networks (communication between them) can disclose individual's brain maturity (thoughts and opinions) through the identification of communicating brain regions, when interpreting stimulus [2]. The accurate discovery of such brain networks can also help the unauthorized people (e.g., attackers) to infer privacy information of an individual. Therefore, it is crucial to focus

on the privacy aspects of the neuroscience research, while developing efficient predictive models, using fMRI data and associated stimulus. In recent years, the concept of compressed sensing (CS) [3] and compressed learning (CL) [4], [5] play major roles in understanding the hidden characteristics of fMRI signals in neuroscience applications and building efficient predictive models.

The compressed sensing (CS) has been developed with the idea of recovering the original signal $x \in X^n$, where x is a vector of n features, from the m measurements $y \in Y^m$ that define m linear relationships between the features and a measurement y with k coefficients that linearize the mathematical relationships, where the condition $m < k < n$ is assumed [6]. The compressed sensing can play a key role in feature selection through the induction of sparsity for removing unnecessary features using the k coefficients. Hence, the dimensionality reduction can be efficiently achieved. Suppose there is an input signal x with features $x = (x_1, x_2, ..., x_n)$, where n is the number of features, along with m number of measurements such that $m < n$. Then we can define the mathematical representation of compressed sensing as follows:

$$y_{m\times 1} = A_{m\times n}(k) \times x_{n\times 1}, \tag{1}$$

where A_k is called the sensing matrix - it is dependent on the k coefficients that can be utilized to generate sparsity for privacy protection and dimensionality reduction.

The compressed Learning (CL) is a mathematical model that incorporates compressed sensing matrix (CS-matrix) with different machine learning algorithms. We get compressed signals from the compressed sensing, and the mechanism of applying different predictive models to that compressed signal to compare the accuracy scores of the model, before and after the use of CS, is the compressed learning. Some of the applications of CL is in compressed image classification, compressive acquisition of dynamic scenes, compressive watermark detection, and compressive hyper-spectral image analysis. Previously, asymptotic analysis of compressed sensing data has been carried out to determine the efficient way to recover the original matrix from the compressed matrix [6]. Similarly, the asymptotic structure in compressed sensing has been studied by Roman et al [7] to better understand the underlying phenomena in practical scenarios and improve results in real-world applications. The paper discussed how effective sampling strategies can be designed through the asymptotic behavior of compressed sensing. In compressed sensing, we can expand and analyze the sensing matrix A to that of A^n, where $n \to \infty$, and asymptotic analysis can be carried over sparse sensing matrices to characterize the compressed sensing model. The sensing matrices used are all sparse. Also, Kotani et al [8] have studied the asymptotic behavior of the transition probability of a random walk on an infinite graph, where they have defined the classical random walk as spatially homogeneous Markov chain, and have characterized the transition probabilities based on the Markov chain.

Our main goal in this paper is to develop a computational framework for the asymptotic-stabilization analysis (as $n \to \infty$) of the compressed sensing model, characterized by transition probabilities, and construct a strong transformation matrix (T_∞) at asymptote to protect the compressed fMRI signals (x) by transforming them. We will also show how we can perform Asymptotically stable analysis based on the

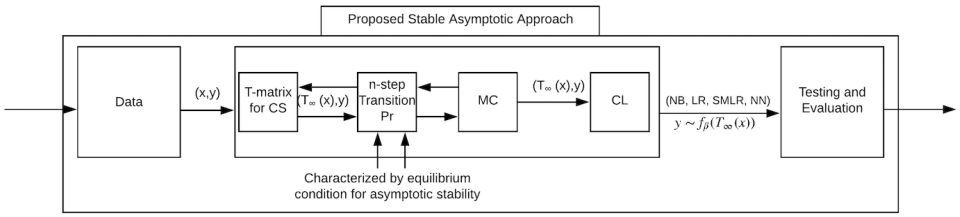


Figure 1: Illustrates the processes of the proposed asymptotic approach. The proposed approach will be evaluated using the NB, LR, SMLR, and NN predictive models.

achieved equilibrium state obtained by calculating iterative n-step transition probabilities and use them to construct a predictive model to protect the privacy of x. The processes involved in our approach can be viewed at Figure 1.

Motivation

In a recent research, we performed a preliminary study on compressed sensing and compressed learning with two-state Markov chain and characterized the transition behavior of fMRI signals [9]. The transition states were used to construct the compressed sensing matrix, which transformed the signals for compressed learning and build a privacy-preserving predictive model. This approach served as a feature selection mechanism. The preliminary results and findings were presented at the Stanford Compression workshop (DOI:10.13140/RG.2.2.16571.87849). We tested the model using its strengths with Logistic Regression (LR) and Sparse Multinomial Logistic Regression (SMLR), while showing significant weaknesses with Naive Bayes (NB) and Artificial Neural Network (NN), which motivated us to perform this research.

There is no one perfect model for any given corpus of data. Rather, there exists many possible models that can be created, each with its own strengths and weaknesses, depending on the scope of the intrinsic patterns and analysis of the data. Whilst, observing this research, it became evident that the one major weakness of this proposed model is that the original fMRI signals (data) which is encrypted by using compressed sensing matrix(A) could very well be decoded by using a brute force attack of trial and error method to deduce the compressed sensing matrix(A). After compressed sensing matrix is deduced, a simple Hadamard product will result in exposing the original fMRI data or signal (x). Let's define it mathematically. A simple parametric model can be presented as follows:

$$y = f_\beta(x) \tag{2}$$

where y is the output (prediction) variable, f defines the predictive model, β is the model parameters, and the vector x is the set of features or in this case the set of original fMRI signals. In our previous work, we proposed a computational approach, where we defined a predictive model, characterized by compressed sensing, as follows:

$$y = f_\beta(A(x)). \tag{3}$$

In this equation, the matrix A is the compressed sensing matrix (CS-matrix) that is computed using two-state Markov chain modeling in our approach. Here, in this

model, if A is deduced by using brute force attack, then the fMRI signal x can be decoded and misused. Thus, this gave the motivation to come up with the asymptotically stable approach to transform the fMRI signals x by using a strong transformation matrix(T_∞) defined by using the same two-state Markov chain by performing iterative computation of n-step transition probabilities to construct a new compressed sensing matrix, such that the signals are protected from brute force attacks. This paper will discuss the above mentioned proposed approach and experimental asymptotic analysis to determine the strong transformation matrix T_∞ to protect the privacy of x whilst comparing the predictive accuracy scores from each analysis.

Methodology

Continuing from the motivation section, The proposed model to overcome the drawbacks of the previous model is represented by using the following parametric model:

$$y = f_\beta(T_\infty(x)) \tag{4}$$

where, T_∞ is the transformation matrix constructed through the computation of n-step transition probabilities, i.e. asymptotic analysis where $(n \to \infty)$. Here, the idea of the proposed model is that, T_∞ can be presented by the n-step evaluation of transition probabilities that is used to determine the compressed sensing matrix A. In addition to that, the transition probabilities chosen should be asymptotically stable for strong T_∞ for the protection of the privacy of the fMRI data. Here, the value of n is crucial to protect the fMRI data, as the change in n will change the transition probabilities which subsequently will change the sequence of x. Furthermore, use of wrong value of n, while decoding will result in different fMRI signal compared to the original fMRI signal. Suppose, x is compressed by using 50-step transition probabilities ($n = 50$), and if the attacker uses $n = 40$ for decoding while performing the brute force attack, then the inverse transformation function (T_∞^{-1}) i.e. $(T_\infty^{-1}T_\infty(x))$ will return an altered or incorrect fMRI signal x; hence, protecting the privacy. The experimental framework of our proposed approach can be viewed in the Figure 1.

We are concerned with ascertaining the asymptotic expression of the transition probabilities which are characterized by a homogeneous Markov chain with state 0 or 1. The transition probability matrix at a particular state may be written by [10]:

$$\begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix} \tag{5}$$

The transition probabilities exhibit behavior ideal for asymptotic analysis. Here, on the basis of asymptotic analysis, we expand the transition probabilities by increasing the size of n such that $(n \to \infty)$, and generate multiple transformation matrix (T_∞) based on value of n. Here, the homogeneous two-state Markov chain also enables us to compute two-step transition probabilities which is given by [10]:

$$\begin{bmatrix} p_{00}^2 + p_{01}p_{10} & p_{01}(p_{00} + p_{11}) \\ p_{10}(p_{00} + p_{11}) & p_{11}^2 + p_{01}p_{10} \end{bmatrix} \tag{6}$$

We can obtain asymptotic expressions for the n-step transition probabilities [10]:

$$\lim_{n\to\infty} p_{00}(n) = \lim_{n\to\infty} p_{10}(n) = \frac{1-p_{11}}{2-p_{00}-p_{11}}, \tag{7}$$

and

$$\lim_{n\to\infty} p_{01}(n) = \lim_{n\to\infty} p_{11}(n) = \frac{1-p_{00}}{2-p_{00}-p_{11}} \tag{8}$$

These equations allow us to analyze the limiting behavior of the transition probabilities under different values of n. Essentially, as we can observe from the asymptotic expressions above, for the expressions to be asymptotically stable, the value of $p_{00}(n)$ should be equal to $p_{10}(n)$, and the value for $p_{01}(n)$ should be equal to $p_{11}(n)$.

Evaluation of Asymptotically Stable Condition for Analysis

There exists a scenario where the value of the transition probabilities will never be equal if we observe all the significant decimal places, as it can be reached to ∞. For example, for 3-step transition probabilities, the value of both $p_{00}(3)$ and $p_{10}(3)$ is 0.015 when rounded off to three-significant decimal places and is given by 3-step computation of transition probabilities, but the value of $p_{00}(3)$ is 0.0150152565 and value of p_{10} is 0.0150251065 when considering ten-significant decimal places for the same 3-step computation, where, $(p_{00}(3) \neq p_{10}(3))$, which should not be the case according to the asymptotic expressions. This basically means that there will be an unstable asymptotic analysis which is unwanted to compute a strong transformation matrix (T_∞). Hence, it is very crucial to decide the asymptotic stability (equilibrium condition) of the model under appropriate small perturbations of initial conditions. We take the significant decimal places of the transition probabilities as the equilibrium conditions under which the model will perform asymptotic stability.

After, computing the $p_{00}(n)$, $p_{10}(n)$, $p_{01}(n)$ and $p_{11}(n)$, we need to define a equilibrium condition, such that when there is a computational domain, and whenever the equilibrium condition belongs to this domain, then the solution will approach to zero as the value of n is increased. Here, the domain is defined by the absolute value of the difference between the transition probabilities. Mathematically, let $p^n(x)$ be the transition probabilities belonging to $p_{00}(n)$ for any size n, and, $q^n(x)$ be the transition probabilities belonging to $p_{10}(n)$ for any size n. Then, the asymptotic stability can be characterized by $|p^n(x) - q^n(x)| \to 0$ as $n \to \infty$. This can also be interpreted as $q^n(x)$ converges to $p^n(x)$ as the size of n increases. So basically, from this we can deduce that if we need to increase the size of n by only small amount, then the model is close to equilibrium condition, while if the required increase in the size of n is relatively large, then the model is more close to the initial condition. The same phenomenon is applied to ascertain the asymptotic stability of $p_{01}(n)$ and $p_{11}(n)$.

fMRI Dataset Origin

In this research, we have utilized the fMRI datasets that are publicly available at the Carnegie Mellon University's StarPlus fMRI data www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www. Some background on the data and the study are discussed in

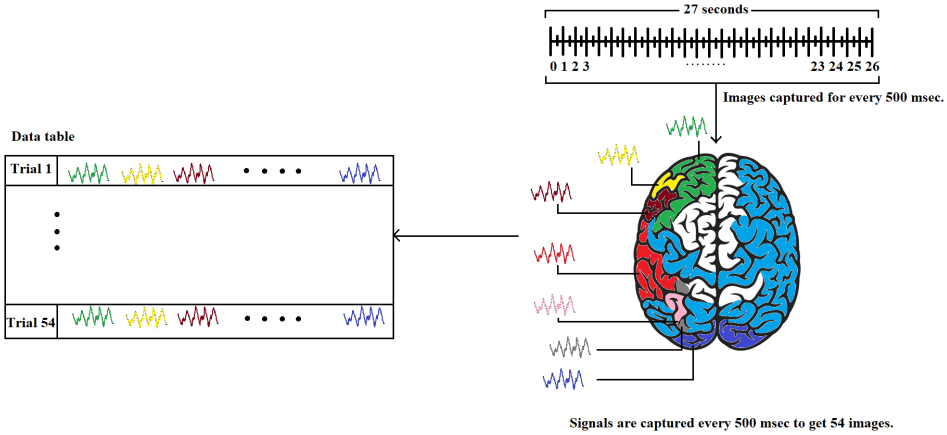


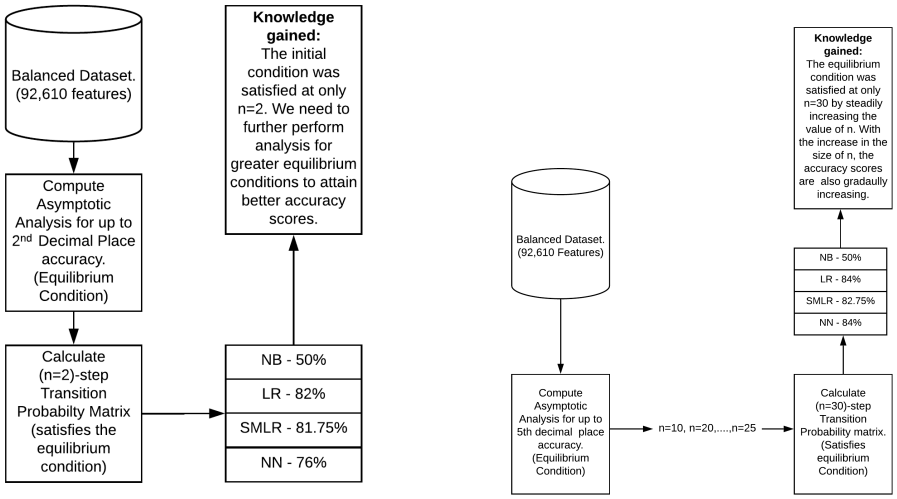
Figure 2: Understanding the dataset by showing the 7 ROIs for subject 04847.

this section. One of the goals of CCBI, as stated at http://www.ccbi.cmu. edu/, was to explain how thought emerges from brain function and if brain dysfunctions show any affect on thoughts. When looking at the data there were systematic steps that were taken with respect to data setup.

The experiment includes a set of trials which lasted for the total of 27 seconds, which was used to create different stimulus to induce different patterns of brain activity. These brain activities were recorded from fMRI scans for every 500 msec. such, that the total observations was given by $(27 \times 2 = 54)$ images. From these images, we could identify different 25 region of interests (ROIs) of brain. And, yet from the experimentation, it is recommended to look at only 7 ROIs of the brain region. From these, the obtained data was shaped for classification. They created a function to analyze the voxels of those ROIs, and, extract the fMRI data signals as points in the voxels belonging to respective ROIs. We performed experimentation on these 7 ROIs which consisted of total 92610 balanced features or signals obtained from the stimulus. The visualization of 7 ROIs is represented in figure 2 which will help to better understand how the fMRI signals was originated. Here, the outline of the top view of the brain is taken form `https://it.clipartlogo.com/istock/brain-top-view-1520441.html`, and is used for creative purpose. Now, we used the fMRI signals(x) that we obtained from this experiment to build a homogeneous two-state Markov chain, which was used to compute transition probabilities, The asymptotic analysis of those transition probabilities will be used to build a predictive model which compresses and changes the original fMRI signals with asymptotic stabilization.

Experimental Analysis

We systematically performed experimentation to understand the effect of asymptotic analysis of transition probabilities while maintaining the asymptotic stability. Note that, the experimentation results are based on the predictive model performance



(Results for 2 decimal place accuracy) (Results for 5 decimal place accuracy)

Figure 3: It presents prediction accuracy scores. Transition probabilities are rounded to the 2^{nd} and 5^{th} decimal places.

upon the balanced fMRI data set. The first step was to determine, the equilibrium condition for the asymptotic stability and compute the n-step probabilities such that, it satisfies the equilibrium condition. The goal was to determine for which value of n, the predictive model will perform the best while protecting the privacy of data. To achieve this, the experimentation was performed on number of domains taking different asymptotic stability conditions into consideration.

Condition for Asymptotic Stability: 2^{nd} *Decimal Place Accuracy*

Firstly, to set a benchmark for our asymptotic analysis, we set an initial condition that the corresponding n-step transition probabilities should be accurate up to 2 decimal places. Then, we started the iterative process of computing the n-step transition probabilities to satisfy the initial condition. The initial condition was achieved in the first iteration itself, i.e. 2-step transition probability. We then used this transition probabilities to construct the transformation matrix for compressed sensing, and applied compressed learning by using the four machine learning models, NB, LR, SMLR and NN. We were able to yield 82% (LR), 80% (SMLR) and 76% (NN), but received only 50% accuracy for NB. The accuracy scores as of now were not great for all the models, especially for NB. The, model performed similarly, for 3^{rd} and 4^{th} decimal place accuracies. The results are presented in Figure 3.

Condition for asymptotic Stability: 5^{th} *Decimal Place Accuracy*

We repeated the process as in stage 1, and performed asymptotic analysis to achieve 5^{th} decimal place accuracy for the corresponding transition probabilities. Then, we performed iterative process to calculate the n-step transition probabilities. First, we calculated the transition probability for size 10 ($n = 10$), and the accuracy was still

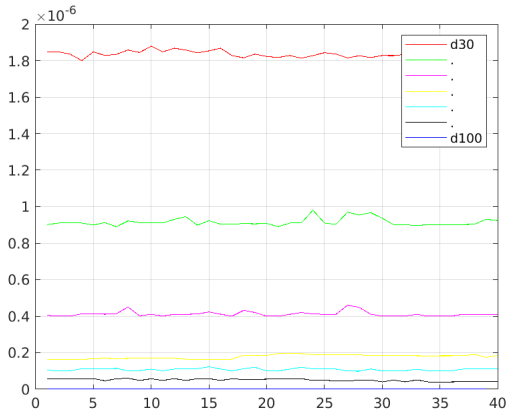


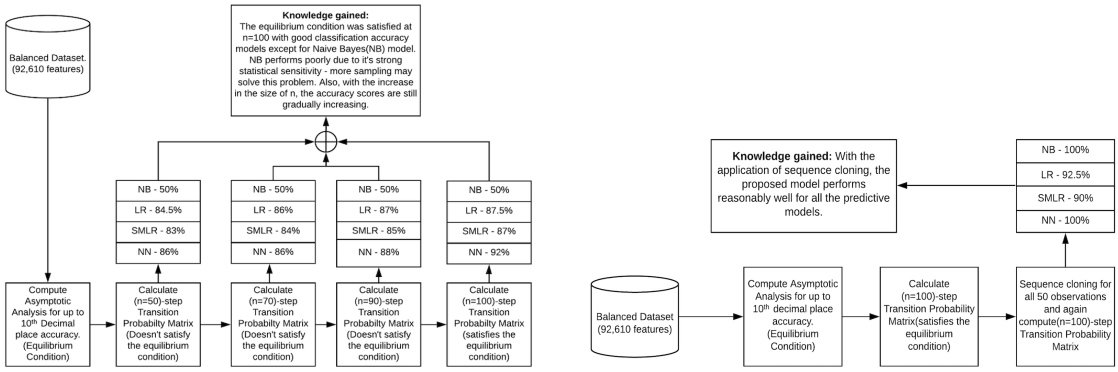
Figure 4: The graph represents the absolute differences between the corresponding transition probabilities($p_{00}(n)$ and $p_{10}(n)$) for the gradual increases of n.

3^{rd} decimal place, hence, we increased the size of n to 20. At $n = 20$, we achieved 4^{th} decimal place accuracy. By further, increasing size of n to 30, we received 5^{th} decimal place accuracy, hence, we reached the equilibrium condition for the asymptotic stability. We once again computed accuracy scores for different models, and, received 84% (LR), 82.75% (SMLR) and 84% (NN), but still the accuracy for (NB) was only 50%. Asides from NB, the accuracy scores for other models were promising, since the increase in size of n to 30 slightly improved the accuracy scores for the models compared to when the size of n was equal to 2.

Condition for the asymptotic Stability: 10^{th} *Decimal Place Accuracy*

Next, after asymptotic analysis at $n = 30$, we increased the required number of accurate decimal places to 10 as the condition for asymptotic stability. Here, in order to represent that we have achieved the asymptotic stability, we calculated the absolute difference of corresponding transition probabilities at different sizes of n, until we got to 10^{th} place decimal accuracy. As discussed in the methodology, as the value of $n \to \infty$, the absolute difference between the transition probabilities belonging to $p_{00}(n)$ and $p_{10}(n)$ should also approach to 0 as we reach the equilibrium condition.

From Figure 4 , we could observe that as we gradually increase the size of n, our absolute difference between the corresponding transition probabilities will gradually start to converge to 0. Here, the absolute difference defines whether the each state of n is near or far from the equilibrium condition(10^{th} decimal place accuracy). As shown in the figure, the value of absolute difference of transition probabilities at $n = 30$ is given by $d_{30} = |p_{00}(30) - p_{10}(30)|$, we can observe that d_{30} is significantly far away from the equilibrium condition, because as we computed earlier, for size $n = 30$, we only achieved up to 5 decimal place accuracy. Hence, when taking 10^{th} decimal place accuracy into consideration, the difference between the two transition probabilities will remain significant. Furthermore, we computed the difference(d) for larger values of n, and observed that the absolute differences is gradually converging to the equilibrium condition as we are achieving more decimal place accuracy. Finally, at $n = 100$, represented by $d_{100} = |p_{00}(100) - p_{10}(100)|$ in the graph, we can see that



(Results for up to 10 decimal place accuracy) (Results for Sequence cloning)

Figure 5: It presents prediction accuracies. Transition probabilities are rounded to the 10^{th} decimal place and sequence cloning is performed for 100-step transition probabilities.

d_{100} converges with 0 i.e. the asymptotic stability has been achieved as we have reached 10^{th} decimal place accuracy. Now, looking at the evaluation of the model, we continued our analysis from $n = 50$, and got the accuracy scores as 50% (NB), 84.5% (LR), 83% (SMLR) and 86% (NN). Then, we increased the size of n, and computed the transition probabilities for $n = 70$ which only had the same values for up to 8^{th} decimal place, and accuracy scores as 50% (NB), 86% (LR), 84% (SMLR) and 86% (NN). Furthermore, we increased the size of n, and evaluated the models for $n = 90$ which gave equal values for up to 9^{th} decimal place, and accuracy scores were 50% (NB), 87% (LR), 85% (SMLR) and 88% (NN). It showed that we were close to achieving the equilibrium condition for asymptotic stability, and a little change in n should be able to achieve equal results for up to the 10^{th} decimal place. We increased the size of n till 100, to make sure that our analysis was asymptotically stable. As shown in figure 4, the absolute difference is equal to 0 in the graph, stressing on the fact that our model was asymptotically stable for up to 10^{th} decimal place accuracy. We got the following accuracy scores, when we used the compressed learning, at $n = 100$, 50%(NB), 87.5% (LR), 87% (SMLR) and 92% (NN). The results are presented in Figure 5. We observed that as the value of n increases, the accuracy of the model is slightly increasing, which characterizes the asymptotic behavior of the given predictive models. From this experimentation, we observed that the Naive Bayes performed poorly, mainly because in Naive Bayes it assumes that the features are independent and doesn't depend upon it's environment. Therefore, it requires multiple samples which are dependent upon each other to estimate the statistical characteristics more accurately in Naive Bayes. As such, we used the sequence cloning technique to generate multiple Markov Chain sequences to define the transformation matrix (T_∞) constructed through asymptotic expressions at different size of n.

Sequence Cloning

We cloned each Markov chain sequences to optimal number of 4, such that, for 50 observations, which resulted in creating the n-step probability matrix with 200 X

92,610 dimensions. Then, as we observed that the predictive model worked best for the larger value of n, we took the value of $n = 100$ and created the transition probability matrix with 200 observations. We tested our approach, using the four machine learning models and saw improvements in every predictive models: 100% (NB), 92.5% (LR), 90% (SMLR), and, 100%(NN). The results are shown in Figure 5.

Conclusions

In this research, asymptotic stability for up to 10 decimal place accuracy was studied, and, evaluated for building the model to transform fMRI signals. However, much more significant asymptotic stability conditions could be considered for further analysis, which would require the transition probabilities to be computed for even larger step size n. Our experiments showed that, as we increase the condition for asymptotic stability, the predictive model performs gradually better, resulting in improved accuracy scores. Finally, based on this study, we concluded that, asymptotic analysis of the transition probabilities computed from homogeneous two-state Markov chain could be used to characterize a strong asymptotically stable predictive model. Hence, this model is capable of transforming the original fMRI signals to a different sets of compressed and asymptotically stabilized fMRI signals for significantly increasing the privacy protection aspect, while providing reasonably good accuracy scores.

References

[1] Shan Suthaharan, *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, Integrated Series in Information Systems. Springer US, 2016.

[2] Suprateek Kundu and Shan Suthaharan, "Privacy-preserving predictive model using factor analysis for neuroscience applications," in *2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity)*. IEEE, 2019, pp. 67–73.

[3] D.L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[4] Amir Adler, Michael Elad, and Michael Zibulevsky, "Compressed Learning: A Deep Neural Network Approach," *arXiv:1610.09615 [cs]*, Oct. 2016, arXiv: 1610.09615.

[5] D.L. Cabaral, "Decoding visual brain states from fMRI using an ensemble of classifiers.," *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 2064–2074, Apr. 2012.

[6] Yaser Eftekhari, A. Banihashemi, and Ioannis Lambadaris, "An efficient approach toward the asymptotic analysis of node-based recovery algorithms in compressed sensing," 01 2010.

[7] Bogdan Roman, Anders Hansen, and Ben Adcock, "On asymptotic structure in compressed sensing," *arXiv preprint arXiv:1406.4178*, 2014.

[8] Motoko Kotani and Toshikazu Sunada, "Asymptotic behavior of the transition probability of a random walk on an infinite graph," *Journal of Functional Analysis*, vol. 159, pp. 664–689, 11 1998.

[9] Michael Ellis, Naseeb Thapaliya, Suprateek Kundu, and Shan Suthaharan, "Illuminating privacy weaknesses in predictive models of f mri data using compressed sensing and compressed learning, DOI:10.13140/RG.2.2.16571.87849," 2019.

[10] Emanuel Parzen, "Stochastic processes," *Holden Day Series in Probability and Statistics, San Francisco: Holden-Day, 1962*, 1962.