# Predicting Human Behavior Based on Survey Response Patterns

# Using Markov and Hidden Markov Models

By

ARUN KUMAR POKHARNA

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

Master of Science in Computer Science

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2016

ACKNOWLEDGEMENTS

Firstly, I would like to express my heartfelt thanks to my supervising professor, Mr. David Levine, who consistently inspired me and motivated me throughout my masters' research. I would also like to thank Dr. Manfred Huber for his continuous guidance and support. Without both of their encouragement and support I could not have imagined this thesis would not have been possible. I am also sincerely Dr. Gergely V. Zaruba for giving valuable suggestions on areas where I overlooked and serving on my committee.

I thank to all my lab mates, especially Sourabh Bose for the brainstorming sessions to come up with the survey designs and other pieces to build the model. I also greatly appreciate the assistance from Azmat, Dr. Vamsikrishna, Bhupender, and other volunteers who helped me in collecting much required survey response data and accommodating my priorities in their busy schedule.

Last but not the least, I would like to thank my family: my grandfather (Bhagwati Lal), my mother(Pushpa), my sister (Monika), brothers (Deepak and Gautam), and all my friends for believing in me and my dreams, for providing me unconditional support, and spreading positivity throughout my years of study. This accomplishment would not have been possible without them. Thank you.

November 21, 2016

ABSTRACT

Predicting Human Behavior Based on Survey Response Patterns Using Markov and Hidden Markov Models

ARUN KUMAR POKHARNA, MS

The University of Texas at Arlington, 2016

Supervising Professor: David Levine

With technological advancements in World Wide Web (www), connecting with people for gathering information has become common. Among several ways, surveys are one of the most commonly used way of collecting information from people. Given a specific objective, multiple surveys are conducted to collect various pieces of information. This collected information, in the form of survey responses, can be categorical values or a descriptive text that represents information regarding the survey question. If additional details regarding the response behavior, scenario in which survey is being responded, or survey outcomes is available, machine learning and prediction modeling can be used to predict these events from the survey data, potentially permitting automatically triggered interventions or preventive actions that can potentially prevent detrimental events or outcomes from occurring.

The proposed approach in this research predicts human behavior based on their responses to various surveys that are administered automatically using an interactive Web–Phone-Computer system. This approach is applied to a typical classroom scenario where students are asked to periodically fill out a questionnaire about their performance before and after class milestones such as exams, projects, and homework. Data collection for this experiment is performed by using Teleherence, a web-phone-computer based survey application. Data collected through Teleherence is then used to learn a predictive model. The approach developed in this research is using clustering to find similarities between different students' responses and a prediction model for their behavior based on Markov and Hidden Markov model.

Contents

List of Figures

List of Tables

# CHAPTER 1

# INTRODUCTION

## 1.1 Introduction

Surveys are among one of the most widely used mechanism of collecting information. There are several ways of conducting surveys known as in-person surveys, web-based surveys, phone call surveys, and text message based surveys. In-person surveys are conducted by interviews, whereas the rest of the surveys do not require a meeting with the candidates. Surveys have been used across many disciplines for various objectives. Several disease studies prefer using survey mechanisms for data collection to conduct research. Khatib et. al. has used a survey mechanism to check medicine availability and affordability [13]. In another study, researchers collected chronic diseases and multi-morbidity data through national health interview surveys [14]. Surveys conducted for a specific objective seek a specific piece of information from the candidates. The objective in this research is to predict human behavior through machine learning and prediction modeling. To observe human behavior, surveys are sent to the students in a classroom. Questions in these surveys attempt to find information about preparation and outcomes for various milestones in a classroom. Milestones are defined as events such as exams, projects, and homework. Responses to the surveys are used as input to clustering algorithm. Results of clustering is used as an input for prediction models such as Markov and Hidden Markov Models. Various experiments are performed to analyze the results of the prediction used here.

## 1.2 Motivation Behind the Thesis

The motivation behind the thesis was to identify the problems that students face at early stages in taking a course such that corrective actions could be taken to improve students' performance in that course and reduce their drop out. If additional details about the students' behavior or outcomes of the milestones are provided, machine learning and prediction modeling can predict these milestones from the survey responses and corrective actions can be taken to intervene and improve the performance of students.

## 1.3 Organization of the Thesis:

The rest of this thesis is organized as follows.

Chapter 2 reviews the related work involving the existing work and other techniques being used in students' performance improvement.

Chapter 3 gives a technical background about Teleherence infrastructure.

Chapter 4 provides a detailed survey design process.

Chapter 5 describes the process of survey response formatting into a vector of vectors.

Chapter 6 proposes the approach and has detailed insights about each segment of the designed algorithm.

Chapter 7 elaborates the experiment setup and analysis of results.

Finally, Chapter 8 discusses the conclusions and possible future work and expansion of this project.

# CHAPTER 2

# RELATED WORK

## 2.1 Existing Methods

In one of the studies, “Survey Analysis System” is designed to analyze the open answers type survey responses referring to automobile brand images. Two tasks are performed in the study: the first task performed is a classification task of assigning the automobile brand image to the car type and the second task was association rules for associating the car type with its answers [10]. In another study, “Association Rule Mining” is applied to questionnaire data [11]. In this study, various data types are identified first and then rule patterns are defined to mine from the questionnaire data. To handle various data types in uniform manner, fuzzy techniques are applied. Through the proposed algorithm, these fuzzy rules are identified. In another study, a CHi-squared Automatic Interaction Detection (CHAID) model based performance prediction model is proposed that identifies the slow learners and study the influence of the dominant factors on their academic performance [12]. This model aims to identify the slow learners and analyze the reasons behind their slow learning process.

## 2.2 Techniques Being Used in Students’ Performance Improvement

Several studies have measured the impact of technology in students’ performance improvement at various levels. At the University of Texas at Arlington, an experiment took place in the summer 2013 for students ranging from second to sixth class where students could use websites, apps, and other resources handy on their devices to teaching handy [17]. Students could use the devices to refer to any material and discuss with the teachers about any issues. At the end of the semester, results of this study were promising as the students’ interest in the class increased and in subsequent semesters use of technological devices also increased. Similarly, in another study conducted among graduate students to find how the students used new technologies and its impact in their learning [18], the students also shared their experience with the technologies in their case studies. Apart from encouraging students to use technologies, in another study, the instructors used technology to teach students and what type of recommendations these instructors gave

parents and caregivers involving technological use [19]. All the above studies directly involve the instructors, the students, or technology used. However, there are indirect ways of conveying the information that can serve as a reminder to students about studying and completing various academic tasks in time. For instance, at times, reminders about the group study after the class may help students join such activities thus resulting in improved performance. Student Veterans Project [9] at the University of Texas at Arlington has been using the Teleherence infrastructure to send reminders, motivational messages, and events notifications to enrolled Student Veterans. Their aim is to acquaint the veterans to academia after their war duties and motivate them about completing education.

# CHAPTER 3

# BACKGROUND

## 3.1 Teleherence

Teleherence is a Web-Phone-Computer based survey application [6]. It is an interactive survey application system where dynamic surveys are designed and scheduled through the phone calls and text messages on the candidate's cell phones. Candidates are the people taking part in a study.

Teleherence provides a technology infrastructure where the candidates information is stored, surveys are designed, scheduled, and sent on candidates' cell phones. The Teleherence infrastructure is mainly divided in three servers TeleDB server, Teleherence server, and Televoicer server. Figure 1: Teleherence Architecture depicts the overall structure of the Teleherence application. Candidates information contains their name, phone numbers, email address, nick name, and pin. This information is stored on the client database server. Since the candidates' data is stored on a separate server, it ensures the data privacy [8] as the care manager cannot find which client they are associated with. They can only connect with the candidates via nick name or pin number.

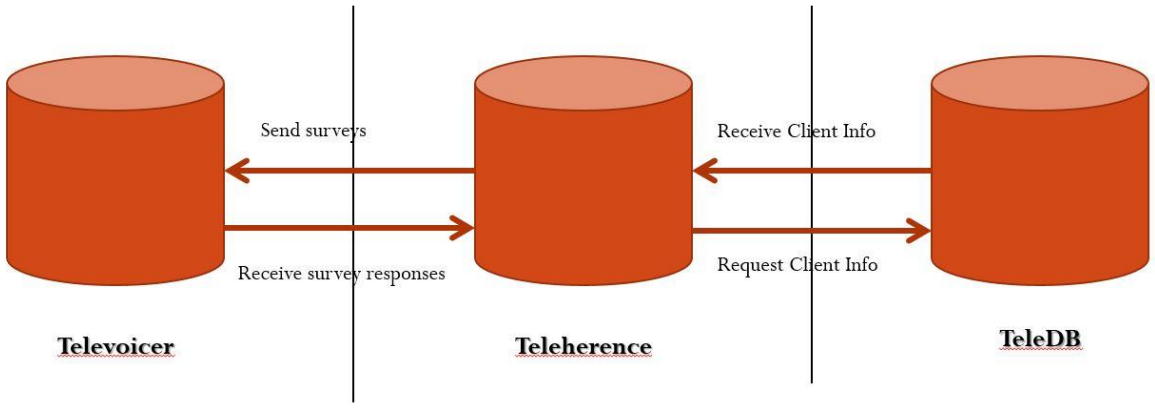


*Figure 1: Teleherence Architecture*

Survey application is stored on the Teleherence server where every segment of a survey is designed. Segments of a survey are “Response Types”, survey items, and templates. Response types are values that would be provided to the candidates as answer choices of survey questions.

All the response choices are stored in a separate table of a Teleherence database, therefore allowing the possibility of reusing a response type in multiple questions. Survey items are questions or information items. Question items provides a piece of information to which the response is desired from the candidates, whereas an information item only provides a piece of information and does not seek any response from the candidates. A question item has a relationship with response types as it needs to define what responses are to be asked along the question. Therefore, a separate relationship table is maintained to store the relationship between the question and associated response types. Both question and information items are stored in a separate table in the Teleherence database, thus allowing the possibility of reusing the same question/information items multiple times in different templates.

A template contains a combination of question and information items. Since it is a sequence of survey items linked in order, it maintains a separate relationship table that stores the order of survey items. If a survey item is a question, it will also have a set of associated response types. Therefore, another table is maintained to store the relationship of templates, items, and response types. Templates are stored in a separate table without being associated with any candidates, therefore reusability of a template is possible for multiple candidates.

When a relationship between a template and a candidate is established, surveys are generated. A separate relationship table is maintained to store templates and candidates. Surveys are created by using a template and associating a scheduled time and recurrence with it to send survey to the candidates. Therefore, another table containing the schedule and recurrence of a survey is stored. A template can be used multiple

times to a single candidate as well as multiple templates can be tied to a single candidate, providing reusability.

A scheduler daemon runs on the Teleherence server that checks the survey schedules every minute. This daemon picks up all the surveys and generate a queue of outstanding surveys. From this queue, the scheduler formats each question/information item of each survey in CCXML form and sends it to the Televoicer server. The Televoicer server is responsible for sending these question/information items on the cell phones of the candidates and waits for their responses in case of question items. Once the responses are received, the Televoicer sends these responses back to the Teleherence server and the results are stored in the results table. Based on the response to first question, an associated next question is sent to Televoicer server. This process continues until all the outstanding surveys are processed. A complete survey structure is not visible to the candidates. Only question/information associated to responses are visible.

This application has been utilized in multiple subjects. The Social work department at the University of Texas at Arlington uses this infrastructure in the Student Veterans Project [9] where they send out surveys and text reminders to veterans regarding various academic activities. This infrastructure has also been used in an anti-smoking study in Taiwan [7]. As of today, December 7th, 2016, this infrastructure is also being used in Autonomous Robotics for Installation and Base Operations (ARIBO). In ARIBO, the Teleherence infrastructure provides a way to schedule rides for an autonomous vehicle and shares the schedule information to other collaborators that handle the autonomous vehicle design.

In the proposed research, another application of the Teleherence infrastructure has been utilized to predict human behavior based on their responses to various surveys. To predict human behavior, students are invited to participate in surveys catering to the information about their preparation and outcomes for various

milestones in a classroom. A brief overview of Teleherence workflow is shown in Figure 2: Overview of Teleherence Workflow.

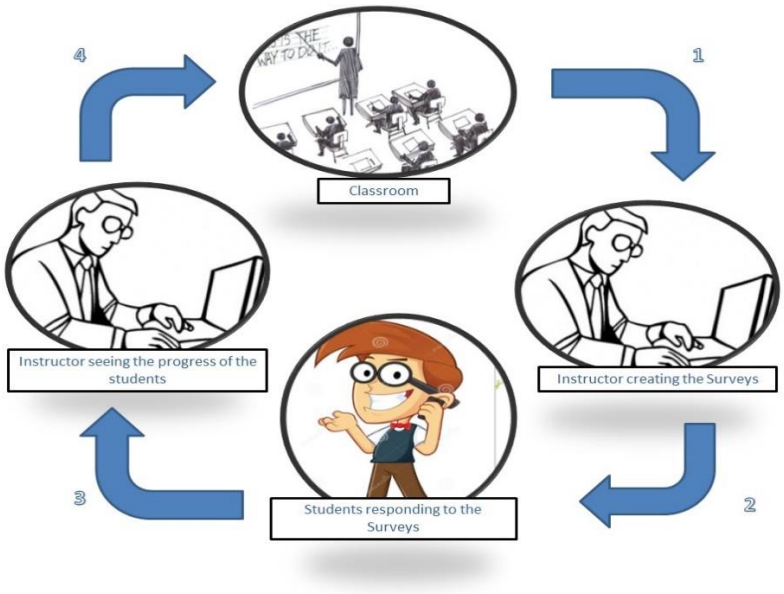


*Figure 2: Overview of Teleherence Workflow*

# CHAPTER 4

## SURVEY DESIGN

Surveys are specifically designed to serve the objective of gathering information from students. To gather this information, surveys are designed in two phases: preparation and outcome. Preparation surveys contains questions seeking information about the preparation of students for a specific milestone, whereas outcome surveys contain questions targeting the responses about the performance of students in that milestone. Milestones are described as different events such as home-work, exam, and projects in a classroom. In this research, 10 milestones are considered from three projects, two exams, and five home-works. Sequence of these milestones is as below:

HW1 – HW2 – Prj1 – Exm1 – HW3 – Prj2 – HW4 – Prj3 – HW5 – Exm2

Since there are two surveys for each milestone, the sequence of surveys for students is:

PreHW1 – PostHW1 – PreHW2 – PostHW2 – PrePrj1 – PostPrj1 – PreExm1 – PostExm1– PreHW3 – PostHW3 – PrePrj2 – PostPrj2 – PreHW4 – PostHW4 – PrePrj3 – PostPrj3 – PreHW5 – PostHW5 – PreExm2 – PostExm2

This order must be maintained because milestones for each student has a fixed order therefore surveys would also attempt to find the same information from the students at each milestone. This order is also important because this study aims to identify if there is an impact of a milestone on other milestones.

Surveys are a directed graph where a question represents vertex and response represents an edge. A vertex can have multiple outgoing edge, representing various responses to the question, and multiple incoming edges representing the responses that leads to this question. The Start vertex does not have any

incoming edge and the end vertex does not have any outgoing edge. An answer to the question will lead to another question in the survey that may differ based on different responses to the question. A complete question-answer sequence is a path traversed from start to end in a graph. This question-answer sequence represents student's response behavior. Questions are asked in the surveys in such a way that they capture various scenarios that a student might face during various milestones. For instance, if a student performed poorly in an exam, the survey would ask questions trying to find if the reason was bad health or missed classes. On the other hand, the preparation survey would ask questions identifying what is impacting his preparations for each milestone.

Surveys for preparation and outcome have different structures. The Preparation survey or pre-survey structure has a standard set of questions for all the milestones. The Start vertex in pre-survey structure asks a question "How is your preparation for the milestone" and has three response edges "Good", "Somewhat good", and "Bad". The Response edge "Good" leads to another vertex that has question as "If needed more material for preparation?". The Response edge "Somewhat good" leads to a different question asking "If outside material is tried". The Question "Facing trouble in understanding material?" is asked when response "Bad" is answered. The Question "If needed more material for preparation?" has two response edges "Yes" and "No". "Yes" leads to another question vertex "Did you talk to professor?", whereas "No" leads to "Do you have high workload?" question. The Question "If outside material is tried" has two response edges "Yes" and "No". "Yes" leads to another question vertex "Did you talk to professor?", whereas "No" leads to "Are you lacking material?" question. The Question "Facing trouble in understanding material?" has response edges "Yes" and "No". When response is "Yes", it leads to next question "Did you talk to professor?", whereas "No" leads to "Are you lacking material?" question. The Question "Did you talk to professor?" has two responses "Yes" and "No" after which survey ends. The Question "Are you lacking material?" has two responses "Yes" and "No" after which survey ends. The Question "Do you have high workload?" has two responses "Yes" and "No" after which survey ends. Figure 3: Pre-Survey Structure shows the structural design of preparation surveys of all the milestones.

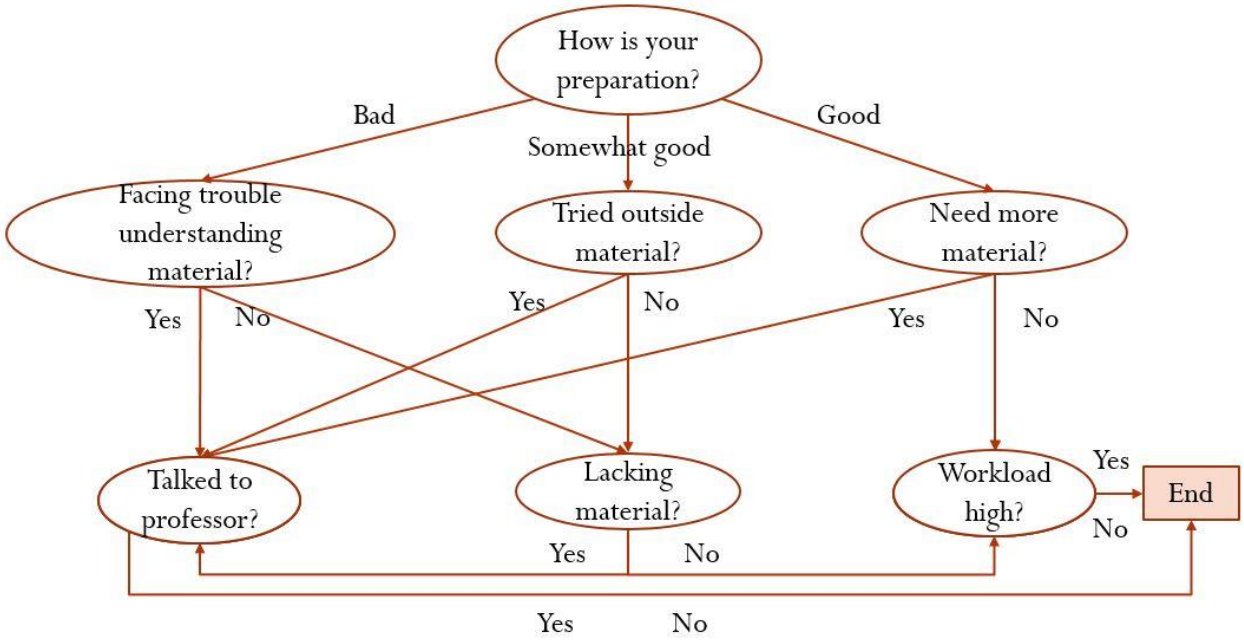


*Figure 3: Pre-Survey Structure*

The outcome survey or post survey structure also has a standard set of questions for all the milestones. The Start vertex in this structure is "How was the milestone outcome?". This question has three response edges "Good", "Somewhat good", and "Bad". All three responses lead to a same question vertex "Better or worse than expectation?" which has two responses "Better" and "Worse". In case of a "Good" response to the first question, the response "Better" leads to end survey, whereas the response "Worse" leads to another question "Problems more complicated than expected?". In the case of "Somewhat good" to the first question, the response "Better" also leads to end survey, whereas the response "Worse" leads to another question "Faced unforeseen circumstance?". In the case of "Bad" to the first question, the response "Better" leads to another question "Was material relevant?", whereas the response "Worse" leads to a different question "Faced any trouble understanding the problems". The question "Problems more complicated than expected?" has two responses "Yes" and "No". "Yes" response leads to end survey, whereas "No" leads to another question "Faced unforeseen circumstance?". The question "Was material relevant?" has two responses "Poorly related" and "Well related". The response "Poorly related" leads to another question

"Talked to the professor?", whereas "Well related" response leads to "Faced unforeseen circumstance?" question. The question "Talked to the professor?" has two responses "Yes" and "No" after which survey ends. The question "Faced unforeseen circumstance?" has two responses "Yes" and "No" after which survey ends. Figure 4: Post-Survey Structure represents the graphical structure of outcome surveys for all the milestones.

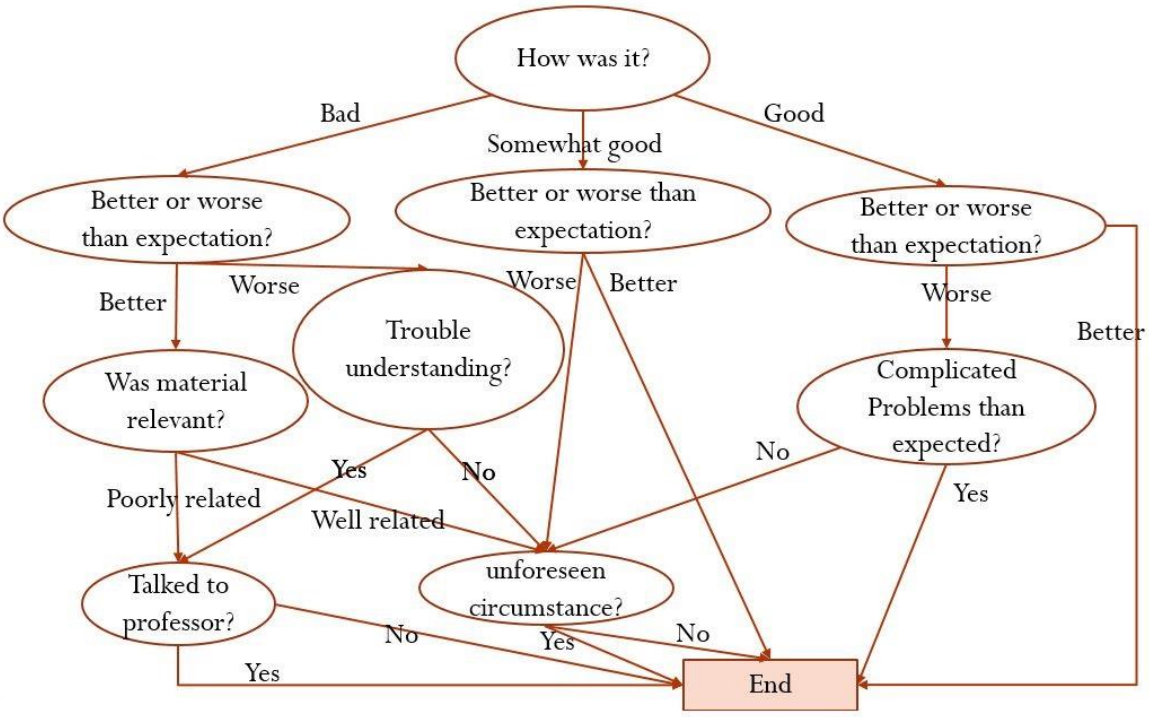


*Figure 4: Post-Survey Structure*

## CHAPTER 5

## RESPONSE STRUCTURE FORMATION

Response edges from the pre-survey and post-survey structures are recorded in a survey_results_tbl of the Teleherence DB. One response edge from the survey is one record of the survey_results_tbl. This table contains the responses to all the survey questions asked to all the students for both the milestones. To identify each row unique to the questions recorded by each student, survey_results_tbl stores result_id (primary key), survey_record_id, template_item_id, response_id, and response_text. survey_record_id represents the outstanding survey for which responses are being recorded. Template_item_id identifies the question in the survey, and response_id identifies the answer provided by the students. The response sequence to one survey is generated by grouping the results on survey_record_id and survey_title from surveys_tbl. After grouping survey responses, a "Response Sequence Vector" is generated separately for pre-survey and post-survey by selecting response_id from the above result. Figure 5: Response Structure Formation illustrates the transition of response_id from table format to a Response Sequence Vector.

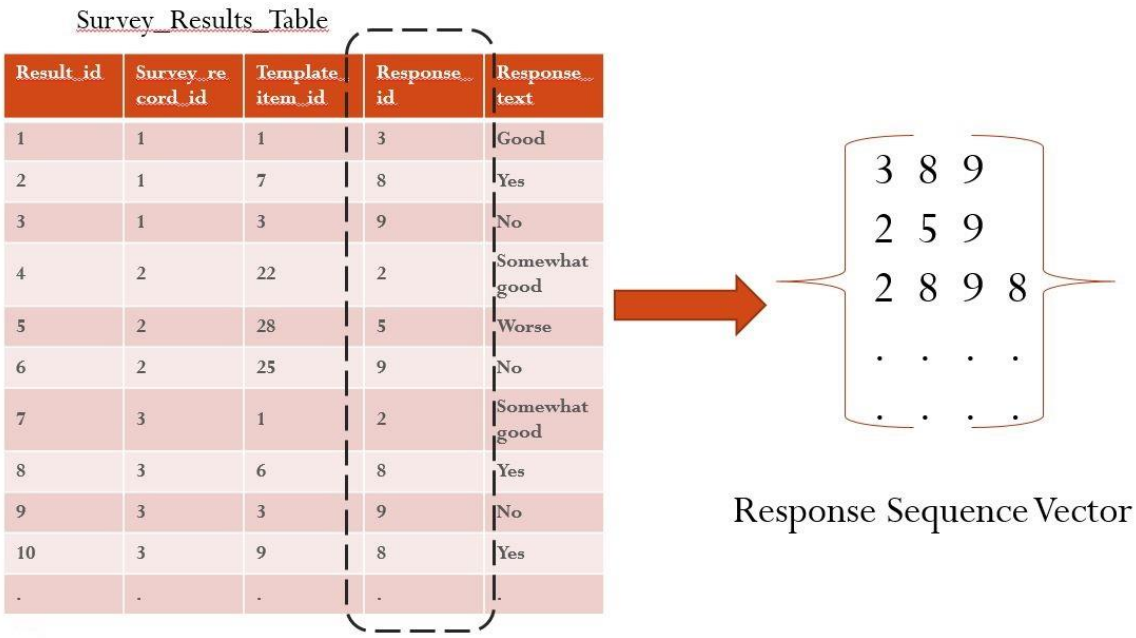


*Figure 5: Response Structure Formation*

## CHAPTER 6

## PROPOSED APPROACH

In this research, a machine learning and Markov model & hidden Markov model based approach is proposed to predict students' behavior through their survey responses. The objective of predicting a students' behavior is to build a model that measures their performance in a classroom so that any corrective actions can be taken at the early stages in the courses that they enroll and help improve their grades by suggesting instructor intervention throughout the semester using survey questionnaire. For instance, if the student faces any trouble in understanding the lectures and it reflects in his survey responses, this model can predict the possible outcomes of every future survey outcomes. Based on this result, the instructor can intervene and take corrective actions that help the student in improving his performance in the course.

The proposed model can predict the future post-survey observation for a student given his pre-survey result for a milestone or predict future pre-survey observations based on current post-survey result. It can also predict the remaining sequence of pre-survey and post-survey results throughout the semester from current results. For example, if the student has responded to a sequence of surveys up to three milestones, this model can predict the rest of survey results for the remaining seven milestones.

In the proposed model, a combination of clustering followed by a Markov/Hidden Markov model is used for prediction. Once array of vectors from the response structure formation process is generated, the pre-survey and post-survey responses are separately used for clustering. Clustering is used to group similar response patterns. After clustering pre-survey and post-surveys, for each student, a sequence of pre-survey and post-survey cluster id for entire semester is generated. An example of such sequence is as below:

{pre1 - post2} - {pre3 - post1} - {pre5 - post 4} - {pre2 - post5} - {pre1 - post2} - {pre4 - post1} - {pre2 - post3} - {pre3 - post2} - {pre4 - post5} - {pre2 - post4}

A pair of "pre" and "post" represents pre-survey which is the preparation survey and post-survey which is outcome survey for a single milestone respectively, and number followed by "pre" and "post" represents cluster id for respective pre-survey and post-survey clusters.

Using the above sequence as input to the Markov Model and Hidden Markov model, prediction for future observations is performed.

## 6.1 Clustering

Clustering is the process of grouping similar patterns in such a way that inter-cluster distance is high and intra cluster distance is low [1]. Clustering is an unsupervised machine learning algorithm where a set of inputs and a set of centroids are provided. In an iterative manner, distance of each input from all given centroids is calculated and minimum distant centroid is assigned as the cluster id for inputs. After each iteration, centroids are recalculated as the mean of all the inputs. Another iteration is run to calculate distance of each point with new centroids and minimum distant centroid is assigned as the cluster id. After each such iteration centroids are recalculated. This process is continued until the new centroids are same as the previous one. This class of machine learning algorithm is also known as Expectation Maximization algorithm.

With the given survey response sequences, clustering helps in grouping students who are responding in a similar fashion. This is an important process because it helps to reduce the overhead of predicting the same results for students responding in a similar way. For instance, if two students have faced similar scenarios during preparation for a milestone and they respond to pre-survey questionnaire in similar fashion, clustering them together will help in time because the result of prediction is going to be same for both these students.

The response sequence array of vectors varies in length based on how students respond to the survey questions. It can have from two to four responses sequences, for pre and post surveys depending on the path in the graph along which students respond. For example, in Figure 6: Sample Path of Responses, if a student responds to sample survey as highlighted, the length of such response sequence is three. Similarly, if the student responds with any other response choice, the resultant length of this vector will be two.

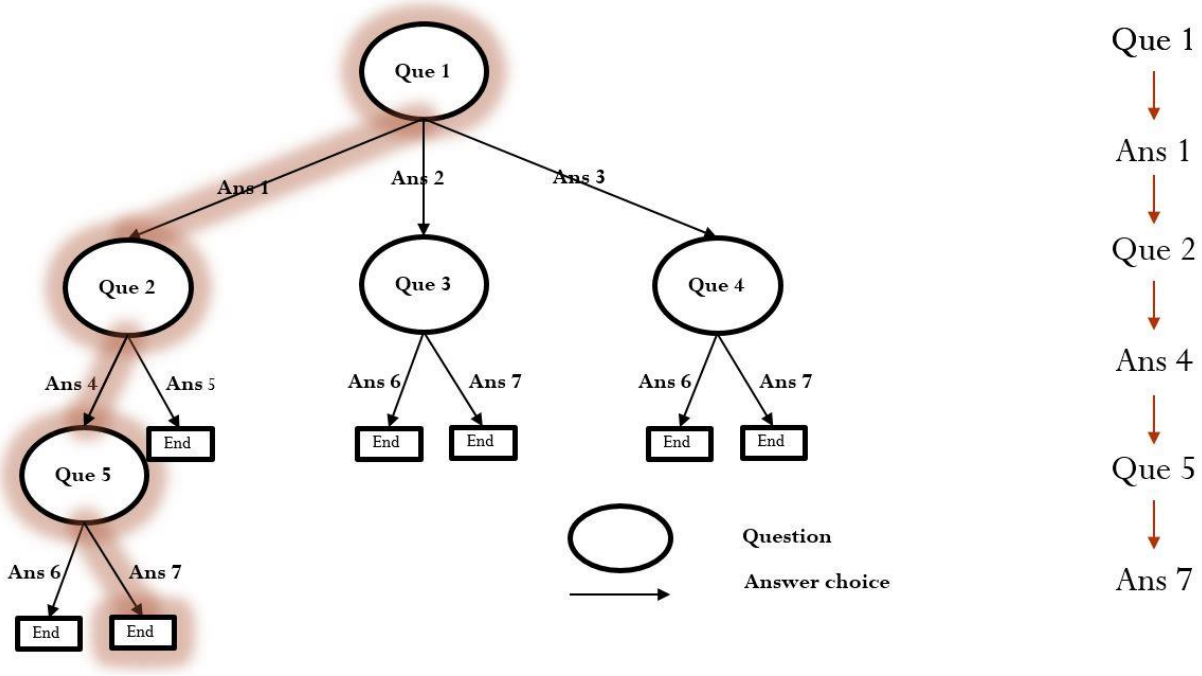


*Figure 6: Sample Path of Responses*

Since input has varying lengths, the centroid of a given cluster cannot be computed using traditional methods. Apart from the varying length response sequences, centroids of the cluster cannot be the mean as it could be a point that does not have a representation from the given set of inputs. Figure 7: Calculating Centroid represents 5 points. Since the circled point has the minimum distance from the rest, it is chosen as the new centroid for this cluster [21].

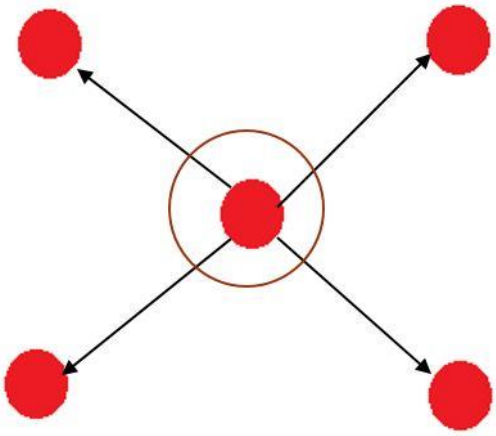

*Figure 7: Calculating Centroid*

Moreover, the responses are categorical values, therefore distance of such input values cannot be calculated in the clustering process by computing the mean of inputs. Therefore, a standard K-means clustering algorithm cannot be used in this unique scenario. To overcome this situation, a customized distance function is needed.

## 6.2 Customized Distance Function

A customized distance function is proposed to calculate distance of an input set of responses with the provided centroids. This customized distance function takes three inputs: centroid, input response vector, and a weight matrix. The weight matrix contains a set of weights. The length of this weight matrix is the length of maximum response sequence from given array of vectors. This weight matrix is used to calculate the distance between provided centroid and input response vector to calculate the distance a comparison in the responses at same level is performed. If the response is same at a given level of centroid and response vector, the result is considered 0. In case of a different response the result is 1. This resultant vector is multiplied with the weight matrix and the result is added to achieve the distance between the

centroid and input response vector. In below example, weight matrix is [5 4 3 2], input is [3 8], and centroid is [6 8 9]. Distance for such input results in 6.

Weight matrix = [5 4 3 2]

Input = [3 8]

Centroid = [6 8 9]

Diff = [1 0 1 0]

Distance = 5 * (1) + 4 * (0) + 3 * (1) + 2 * (0) = 6

## 6.3 Weight Learning Overview

The weight matrix used above determined based on prior knowledge of the surveys. Therefore, it relies highly on human insights. However, these weights can also be learned. To learn these weights, we need to optimize some function which can maximize the accuracy of the model or maximize some other user defined reward.

Finally, to summarize the clustering process, inputs are number of clusters and response sequence array of vectors as samples for clustering. Initially, centroids are randomly assigned for the given number of clusters. A loop is run until the new assigned clusters are the same as previous ones. In this loop, distance for each sample is calculated with all the centroids. Cluster id is assigned to the sample having minimum distance. Once all the inputs are assigned a cluster id, new centroids having minimum distance with all the other points in the cluster are calculated for all the clusters.

Clustering is performed for pre-survey and post-survey response sequence array of vectors separately and respective cluster ids are calculated for each survey. After clustering, cluster id sequences are generated following the sequence in which the surveys were sent. Figure 8: Cluster id Sequence Formation from Response Sequence Vector shows the transformation of response sequence vector into cluster id sequences as an outcome of clustering.

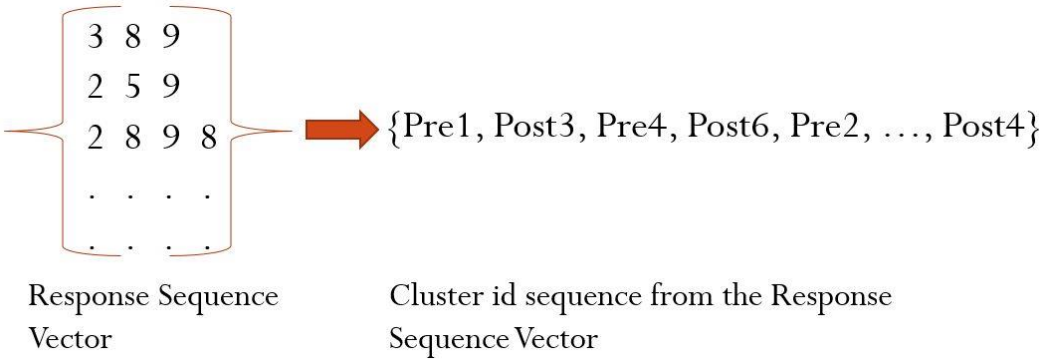


*Figure 8: Cluster id Sequence Formation from Response Sequence Vector*

### 6.4 Markov Model

A Markov model is a stochastic model used to model randomly changing systems where it assumes the Markov Property. A Markov property is when the future states depend only on the current state not on the events that occurred before it [5].

The Markov Model is trained from the cluster id sequences generated at the end of clustering process. To train Markov Model, 70% of data is used. In the training process, the transition matrix is created by observing the transitions from pre-cluster id to post-cluster id and post-cluster id to pre-cluster id. Feeding sequences of all the students in the training data for the entire semester gives us the transition probability matrix that represents the probability of an outcome given preparation, for example, what will be an outcome

for an exam if the preparation was good. Similarly, the quality of preparation of an upcoming milestone can be computed given the outcome of previous milestone.

## 6.5 Training Data for Markov Model

Training data is a vector of vectors with fixed length of 20. Each row represents a student and columns represents all the 20 pre and post cluster id sequences. A sample data for cluster id sequence is shown below.

{pre1 - post2} - {pre3 - post1} - {pre5 - post4} - {pre2 - post5} - {pre1 - post2} - {pre4 - post1} - {pre2 - post3} - {pre3 - post2} - {pre4 - post5} - {pre2 - post4}

{pre3 - post1} - {pre1 - post2} - {pre5 - post4} - {pre3 - post2} - {pre1 - post1} - {pre2 - post5} - {pre5 - post1} - {pre2 - post3} - {pre4 - post4} - {pre1 - post2}

Prediction for future observations sequence is done using a transition probability matrix from the trained model. For instance, a student has responded to the surveys for first 6 milestones i.e. 12 pre and post survey questionnaire. Observations for the remaining 4 milestones can be predicted using Markov Model. For example, above partially observed sequence for 6 milestones is as below:

{pre3 - post1} - {pre1 - post2} - {pre5 - post4} - {pre3 - post2} - {pre1 - post1} - {pre2 - post5}

Since a Markov model assumes that future observations depend only on current state of Markov Model [4], pre-cluster id for 7th milestone would only depend on post cluster id i.e. "post5" in above example.

## 6.6 Proposed Markov Model Design

A Markov Model designed for the proposed research is comprised of 3 variable state space. Thus, each state contains 3 cluster ids as (postX1, preX2, postX3). Here X1, X2, and X3 represents respective cluster ids. Since Start of the model does not have a postX, therefore, a special variable "Start" is used. Each future observation is predicted as a pair of pre and post cluster ids for the milestone. Observation Sequence for Pre1, Post2, Pre2, Post3, Pre4, Post4, Pre3, Post2 is highlighted in Figure 9: Proposed Markov Model.

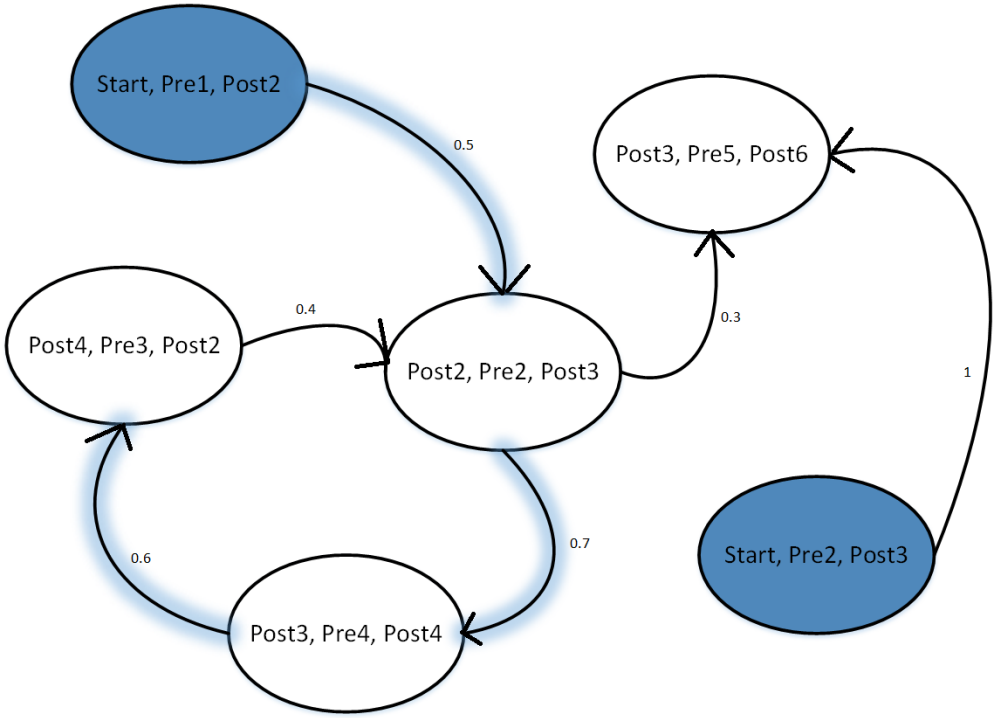


*Figure 9: Proposed Markov Model*

## 6.7 Hidden Markov Model (HMM)

A Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states [2]. HMM is represented as the output states are not visible but the observations dependent on the state are visible. Therefore, the prediction will not only be dependent on the transition probability but emission probability. Emission probability is the probability of

emitting an observation Y given state X. Figure 10: Proposed Hidden Markov Model represents the proposed HMM for this research.

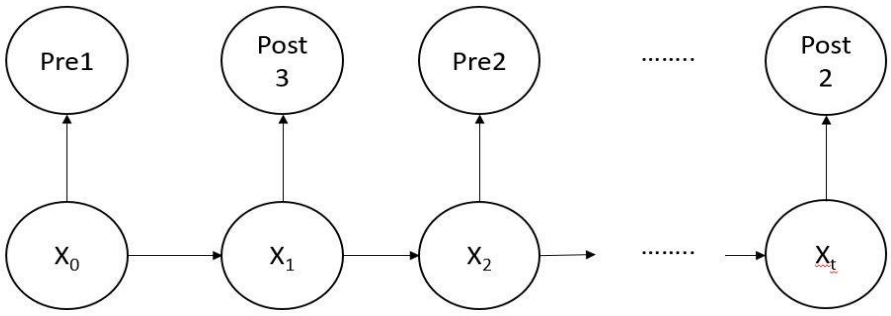


*Figure 10: Proposed Hidden Markov Model*

## 6.8 Prediction using HMM

In the proposed research, observations are represented as pre and post cluster ids. States model the underlying relationship between the observations variables, information about which is not present in the data. In the proposed model, states do not have a definition, therefore, by looking at the complexity of the data, numerous states are assumed and future observations are predicted.

## 6.9 Training Data for Hidden Markov Model

Training data for HMM is the same as that of a Markov Model, i.e. pre and post cluster id sequences for all the milestones for all the students. Like Markov Model, 70% of the data is used for training the Model. Baum-Welch algorithm [3] is used to train the HMM.

HMM assumes that future observations depend on all the previous states as well as current state. For instance, in the below partially observed sequence of six milestones, future observations will depend not only on the state of post5 observation but all previously observed states of the observation sequence.

{pre3 - post1} - {pre1 - post2} - {pre5 - post4} - {pre3 - post2} - {pre1 - post1} - {pre2 - post5}

## 6.10 Simulated Annealing

Simulated Annealing(SA) is an optimization process to achieve a global optimum of a given function [15]. In the proposed approach, SA is used to optimize weights for clustering. Input for SA are Markov Model and Hidden Markov Model. SA assigns random weights in a weight matrix and run Markov Model and Hidden Markov Model in a loop. After every iteration, the accuracy is compared from the previously computed accuracy and weight matrix with optimized accuracy is considered. Exit condition for this loop is when exit criteria is matched. Exit criteria is the cooling parameter reduced to zero. At the end, a set of learned weight matrix is achieved and this weight matrix is used for clustering and prediction.

# CHAPTER 7

# EXPERIMENTS AND RESULTS

## 7.1 Experimental Setup

Input to the proposed model is a Response Sequence Vector which is generated by formatting survey_results_tbl of Teleherence DB. To perform the research, a classroom environment is emulated where five homework, three projects, and two exams were considered as milestones. Each milestone has two surveys pre-milestone survey and post-milestone survey. Therefore, each student responds to 20 surveys to complete a survey sequence for the semester. Every student assumes a scenario while responding the survey questionnaire. A scenario is defined as a situation under which the student responds to the surveys for any milestone. Table 1: Considered Scenarios for each Milestone shows various scenarios considered for data collection for different milestones in the entire semester.

| Events | Scenario1 | Scenario2 | Scenario3 | Scenario4 | Scenario5 |
|---|---|---|---|---|---|
| **HW1** | Good | Good | Good | Missed class | Good |
| **HW2** | Good | Good | Good | Missed class | Good |
| **Prj1** | Good | Good | Good | Good | Good |
| **HW3** | Fallen Sick | Good | Good | Good | Missed class |
| **Exm1** | Fallen Sick | Good | Good | Good | Good |
| **Prj2** | Good | Good | Good | Fallen Sick | Good |
| **HW4** | Missed class | Good | Missed class | Fallen Sick | Good |
| **Prj3** | Good | Good | Fallen Sick | Good | Fallen Sick |
| **HW5** | Good | Good | Fallen Sick | Good | Good |
| **Exm2** | Good | Good | Good | Good | Good |

*Table 1: Considered Scenarios for each Milestone*

Clusters are trained on 70% of the data, therefore, 24 students response sequences are used as training data. Validation dataset is 15% of total data i.e. six students' response sequences. The validation dataset is used to improve validation accuracy in Simulated Annealing process. The remaining 15% i.e. five students' response sequences are used as test dataset.

## 7.2 Proposed Approach for Predefined Weight Matrix

Response Sequence Vectors are considered as an input to the modified K-means Clustering phase. The weight matrix in this phase is constant and used as [4 3 2 1]. The result of clustering phase is a set of pre-cluster and post-cluster id sequences. This input sequence is used as input to both the Markov and Hidden Markov Model. Output of Markov and Hidden Markov Model is the prediction result. Figure 11: Proposed Approach for Predefined Weight Matrix represents the flow diagram for proposed approach for Predefined Weight Matrix.

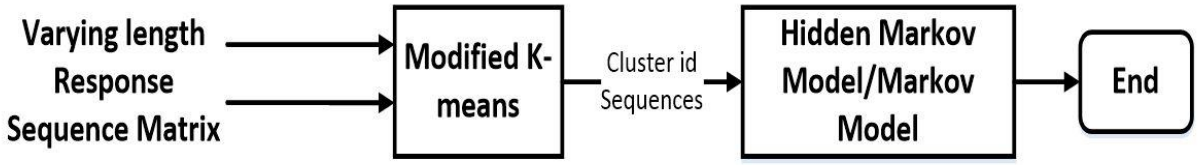


*Figure 11: Proposed Approach for Predefined Weight Matrix*

## 7.3 Proposed Approach for Predefined Weight Matrix

Unlike in **Error! Reference source not found.**, Response Sequence Vectors along with randomly initialized weight matrix are considered as an input to the modified K-means Clustering phase. Outcome of

this phase is a set of pre-survey and post survey cluster-id sequences. This input sequence is used as input to both the Markov and Hidden Markov Model. Output of Markov and Hidden Markov Model is the prediction result. However, Prediction Accuracy Validation is done in Simulated Annealing phase and every time a new weight matrix is generated randomly. Results of this process keep staying in the loop until (Accuracy X # clusters) is achieved maximum. After receiving the learned weight matrix test data is used to perform the prediction of future observations. Figure 12: Proposed Approach for Learned Weight Matrix represents the flow diagram for proposed approach for the Learned Weight Matrix.

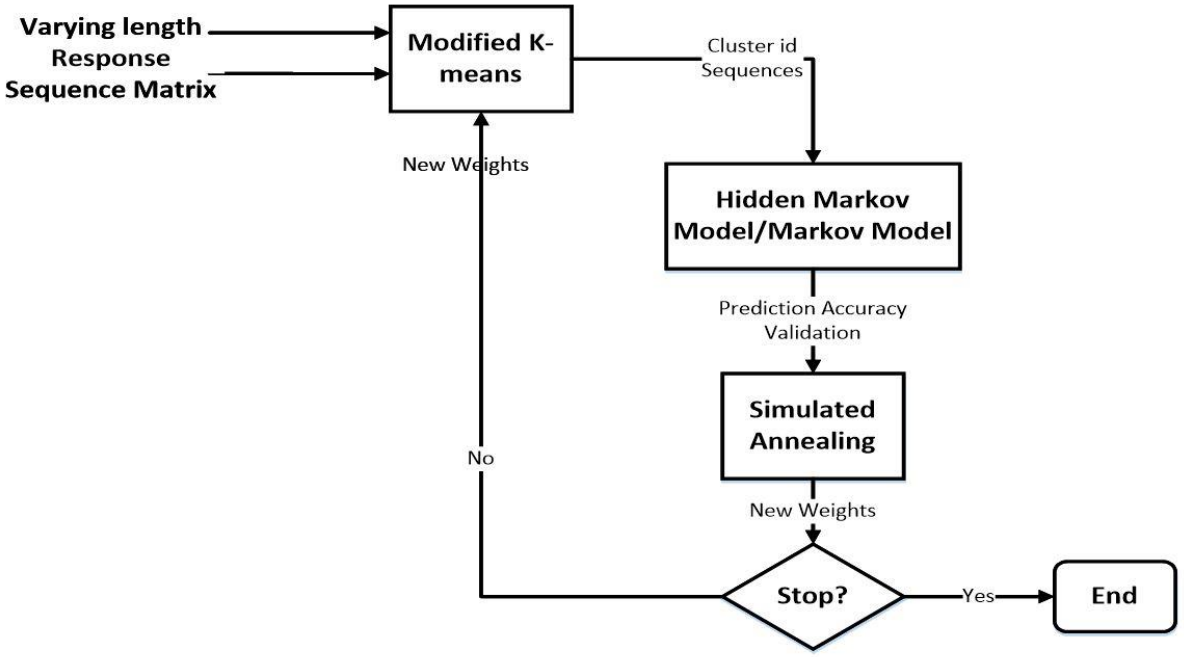


*Figure 12: Proposed Approach for Learned Weight Matrix*

## 7.4 Results for Markov Model with Predefined Weights

Results achieved in this experiment are based on a predefined weight matrix [4 3 2 1]. Figure 13: Markov Model Based Prediction Using Predefined WeightsFigure 13: Markov Model Based Prediction Using Predefined Weights represents the accuracy plot for given number of clusters. X-axis represents the given number of clusters and remaining clusters are predicted with the accuracy measured on Y-axis.

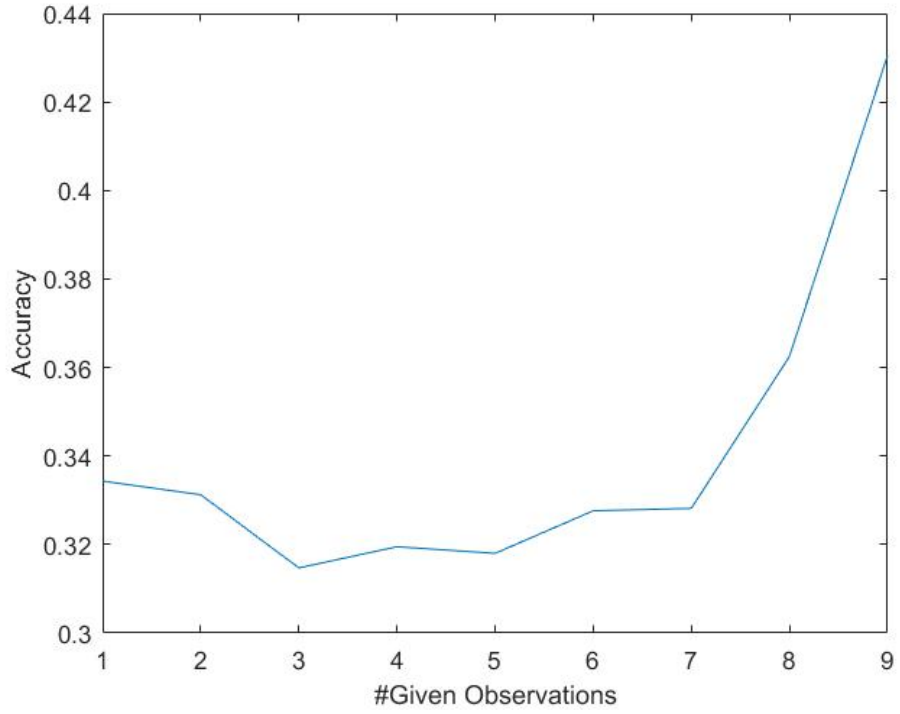


*Figure 13: Markov Model Based Prediction Using Predefined Weights*

## 7.5 Results for Hidden Markov Model with Predefined Weights

Results achieved in this experiment are based on a predefined weight matrix [4 3 2 1]. From the Figure 14: Accuracy for #Given Observations Vs #States for Initial #States= 10, Jump = 10 states, End state = 150, it is clearly visible that prediction accuracy is high when number of states for prediction are smaller. Also, for 80 states, the prediction accuracy is much better across all given observations. When the number of states are increased to more than 80, prediction for pre-cluster id decreases consistently.

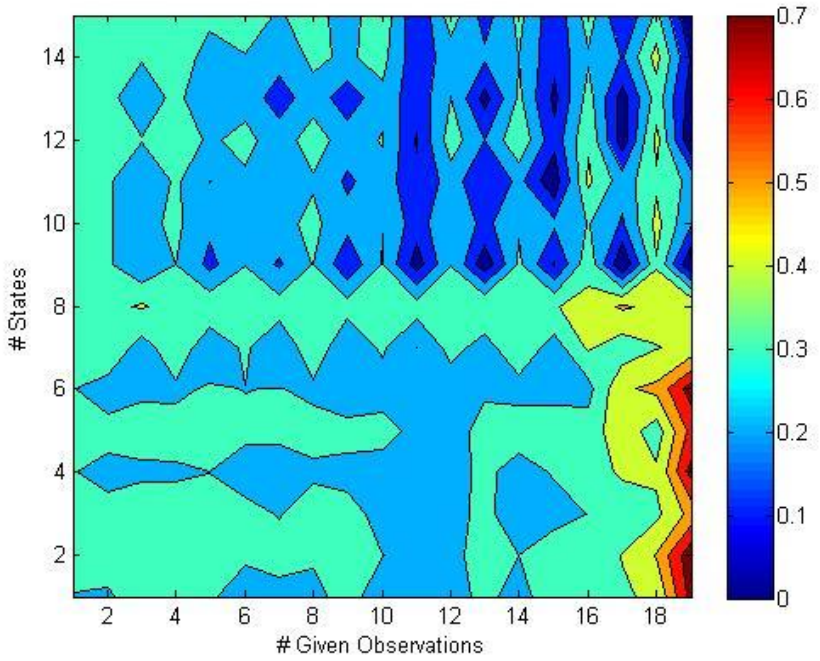


*Figure 14: Accuracy for #Given Observations Vs #States for Initial #States= 10, Jump = 10 states, End state = 150*

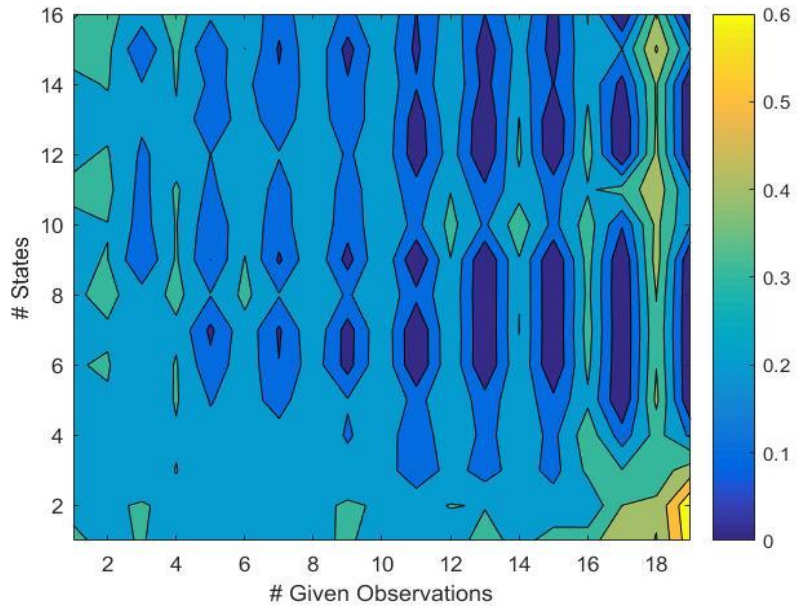


*Figure 15: Accuracy for #Surveys to be Predicted Vs #States for Initial #states= 50, Jump = 20 states, End state = 370*

Figure 15: Accuracy for #Surveys to be Predicted Vs #States for Initial #states= 50, Jump = 20 states, End state = 370

represents another experiment where initial states are 50 and after every iteration for given observations 20 states are increased. Iteration continues until the states reaches to 370. The contour graph shows that this model shows a tendency to overfit when there are less observations to predict for given states between 50 to 70 states. The accuracy is highest in that region. Moreover, the pre-survey cluster id prediction is below 10% when given observations are 10 or more.

## 7.6 Results for Markov Model with Learned Weights

Results achieved in this experiment are based on a learned weight matrix using Simulated Annealing process. Optimized weight matrix achieved through Simulated Annealing on Markov Model is [0.732532703600342 0.773076478524415 0.035180251199843 0.968583021731241]. Simulated Annealing results are shown inFigure 16: Simulated Annealing for Markov Model.

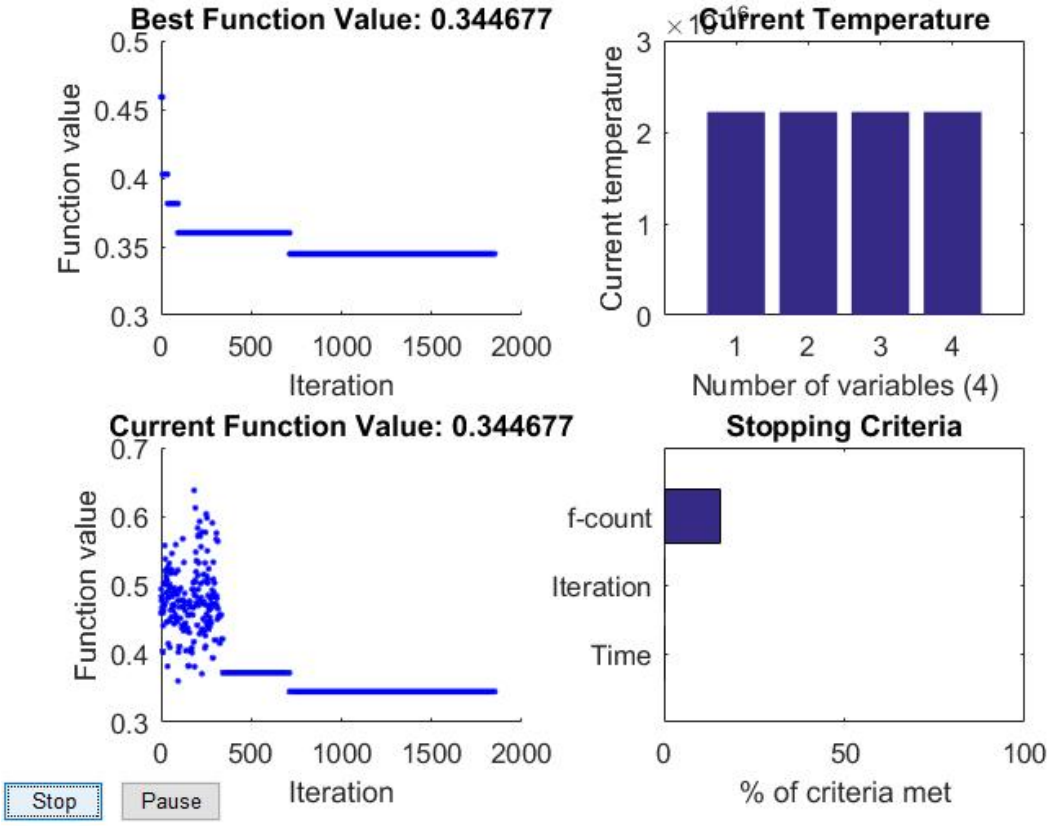


*Figure 16: Simulated Annealing for Markov Model*

Figure 17: Markov Model Based Results Using Learned WeightsFigure 17: Markov Model Based Results Using Learned Weights represents the accuracy plot for given number of clusters. X-axis represents the given number of clusters and remaining clusters are predicted with the accuracy measured on Y-axis.

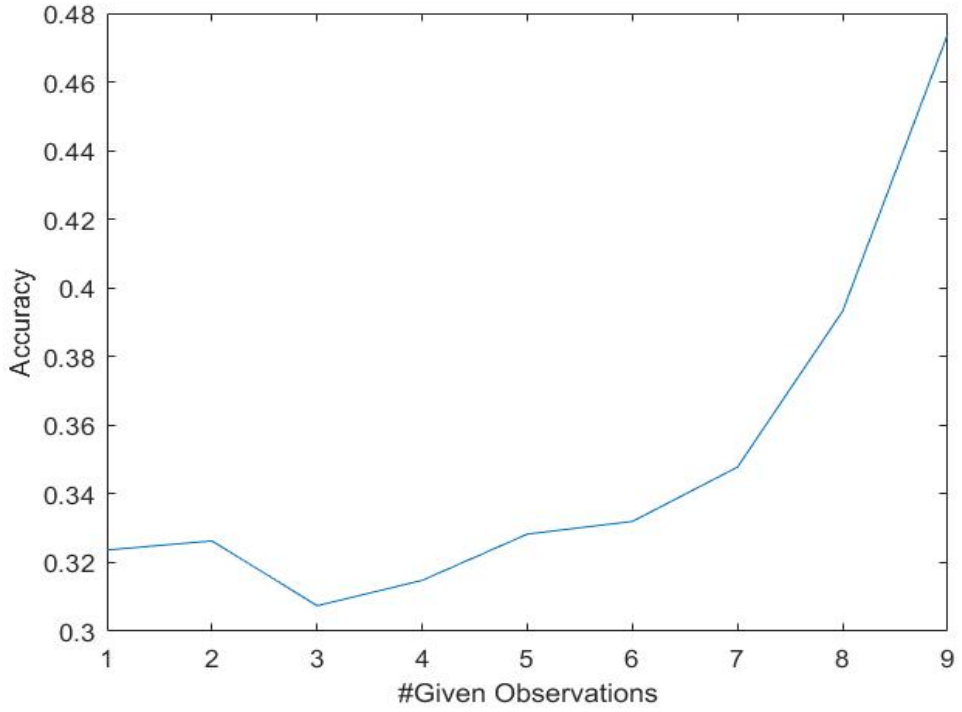


*Figure 17: Markov Model Based Results Using Learned Weights*

## 7.7 Results for Hidden Markov Model with Learned Weights

Results achieved in this experiment are based on learned weight matrix using Simulated Annealing process on Markov Model. The learned weight matrix achieved is [0.732532703600342 0.773076478524415 0.035180251199843 0.968583021731241]. From the Figure 18: Accuracy for #Given Observations Vs #States for Initial #States= 10, Jump = 10 states, End state = 150, it is clearly visible that prediction accuracy is high when number of states for prediction are low. Also, for states between 70 to 80, the prediction accuracy is much better across all given observations.

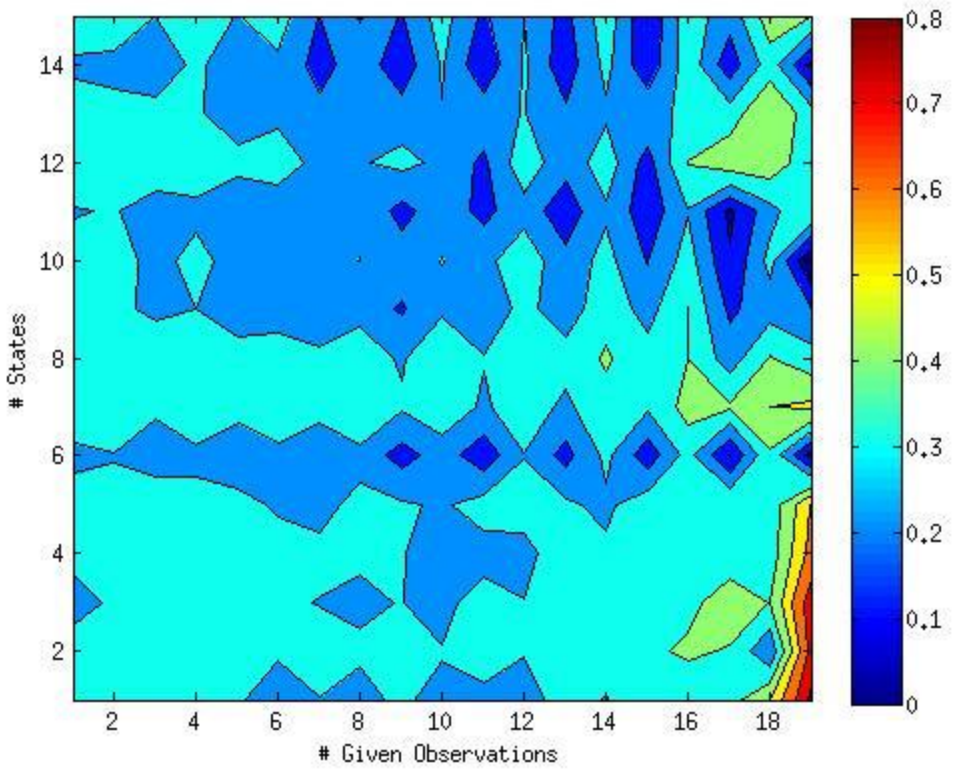


*Figure 18: Accuracy for #Given Observations Vs #States for Initial #States= 10, Jump = 10 states, End state = 150*

# CHAPTER 8

# CONCLUSION AND FUTURE WORK

## 8.1 Conclusion

In this research, the proposed approach extracts survey responses from the Teleherence DB and clusters them. Subsequently, these cluster sequences are then used to model sequences of responses using both Markov and Hidden Markov Model for comparison. If the Markov assumption is made, the Markov Model is used to predict a combination of pre-survey and post-survey clusters. Similarly, in case of Hidden Markov Model, pre and post observations are also predicted but the future observation is dependent on the entire previous observation sequence.

The experiments show comparable results between the Markov and Hidden Markov Model results, however, Hidden Markov Model seems to provide slightly better test accuracy.

## 8.2 Future Work

The proposed model was built on simulated data. Thus, as a possible future work, an actual dataset from real world classes with ongoing milestones need to be collected to measure the performance of the proposed approach on real world data.

Additionally, since Simulated Annealing is extremely computationally expensive, learning the weights without requiring the Markov Model and Hidden Markov Models is needed. To address this problem, a possible future work could be to learn weights based on a system of rewards which adhere to the specific survey needs.

In the current design, the dataset used is inherently ambiguous. Since a few survey questions has negation relations between them and their response types are the same, therefore, even though their meaning is different, they are grouped together which is ambiguous. To remove such misinformation static representation of the responses is required. Therefore, more tests are needed to compare model performances using the existing method with varying response sequences and the one with added unanswered question variables in the existing response sequence thus making it static.

## REFERENCES


[1] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

[2] Blunsom, Phil. "Hidden markov models." *Lecture notes, August* 15 (2004): 18-19.

[3] Welch, Lloyd R. "Hidden Markov models and the Baum-Welch algorithm." *IEEE Information Theory Society Newsletter* 53.4 (2003): 10-13.

[4] Klebaner, Fima C. *Introduction to stochastic calculus with applications*. Vol. 57. London: Imperial college press, 2005.

[5] Atkinson, Richard C., and Patrick Suppes. *An analysis of a two-person interaction situation in terms of a Markov process*. Applied Mathematics and Statistics Laboratory, Stanford University, 1957.

[6] Ramachandran, Divya, et al. "Technology to Help Patients Adhere to Treatment Plans–A Brief Introduction to the Teleherence Project." *Proceedings of the Biotechnology and Bioinformatics Symposium (BIOT-2008), Arlington, Texas*. 2008.

[7] Peng, Wu-der Brian, and Dick Schoech. "Evaluation of a web-phone intervention system in changing smoking behavior—A randomized controlled trial." *Journal of Technology in Human Services* 31.3 (2013): 248-268.

[8] Schoech, Dick, and Kristin Whitehill Bolton. "Automating and Supporting Care Management Using Web-phone Technology: Results of the 5-Year Teleherence Project." *Journal of Technology in Human Services* 33.1 (2015): 16-37.

[9] Smith-Osborne, Alexa M. "Supporting resilience in the academic setting for student soldiers and veterans as an aspect of community reintegration: The design of the Student Veteran Project study." *Advances in Social Work* 13.1 (2012): 34-50.

[10] Yamanishi, Kenji, and Hang Li. "Mining open answers in questionnaire data." *IEEE Intelligent Systems* 17.5 (2002): 58-63.

[11] Chen, Yen-Liang, and Cheng-Hsiung Weng. "Mining fuzzy association rules from questionnaire data." *Knowledge-Based Systems* 22.1 (2009): 46-56.

[12] Ramaswami, M., and R. Bhaskaran. "A CHAID based performance prediction model in educational data mining." *arXiv preprint arXiv:1002.1144* (2010).

[13] Khatib, Rasha, et al. "Availability and affordability of cardiovascular disease medicines and their effect on use in high-income, middle-income, and low-income countries: an analysis of the PURE study data." *The Lancet*387.10013 (2016): 61-69.

[14] van Oostrom, Sandra H., et al. "Time Trends in Prevalence of Chronic Diseases and Multimorbidity Not Only due to Aging: Data from General Practices and Health Surveys." *PLoS One* 11.8 (2016): e0160264.

[15] Kirkpatrick, Scott. "Optimization by simulated annealing: Quantitative studies." *Journal of statistical physics* 34.5-6 (1984): 975-986.

[16] Müller, Meinard. "Dynamic time warping." *Information retrieval for music and motion* (2007): 69-84.

[17] Arrowood, Dana. "Bring Your Own Device: Using What You Have in a Preservice Teacher Preparation Class." *Society for Information Technology & Teacher Education International Conference*. Vol. 2014. No. 1. 2014.

[18] Davis, Ruth, and Dana Arrowood. "Inservice Teachers Self-Report of Technology Competence: Effects of Implementation of Technology Activities During a Case Study." *Society for Information Technology & Teacher Education International Conference*. Vol. 2012. No. 1. 2012.

[19] Arrowood, Dana, et al. "Supporting preservice teachers as they use technology to teach children." *Society for Information Technology & Teacher Education International Conference*. Vol. 2010. No. 1. 2010.

[20] Arrowood, Dana. "The Classroom FabLab Project."

[21] Loganathan, Ashokkumaar P., and Manfred Huber. "An approach for behavior discovery using clustering of dynamics." *Systems, Man and Cybernetics, 2008. SMC 2008. IEEE International Conference on*. IEEE, 2008.

## BIBLIOGRAPHICAL STATEMENT

Arun Kumar Pokharna received his Bachelors of Engineering degree in Information Technology from Manikya Lal Verma Textile and Engineering College, Bhilwara under the University of Rajasthan, Jaipur, Rajasthan, India in 2009.

After his graduation, he worked in Accenture Services Pvt. Ltd for 3.5 years in Mumbai, India. In the United States, he worked as a Graduate Research Assistant at the University of Texas at Arlington and has been working on various applications of Teleherence.

His current research interests are Data Analysis, Prediction Modeling, and applied Machine Learning.