Application of Hidden Markov Modeling to Objective Medical Skill Evaluation

Thomas MACKEL [1], Jacob ROSEN [1,2], Ph.D., Carla PUGH [3], M.D., Ph.D.
[1] *Department of Electrical Engineering,* [2] *Department of Surgery, University of Washington, Seattle, WA, USA*
[3] *Department of Surgery, Northwestern University, Chicago, IL, USA*
E-mail:{tmackel ,rosen}@u.washington.edu drpugh@northwestern.edu
Biorobotics Lab: http://brl.ee.washington.edu

Abstract: A methodology using Hidden Markov Modeling (HMM) is used to analyze medical procedural data from the E-Pelvis database. The focus is on the method of selection of HMM parameters. K-Means is used to choose the alphabet size. Successful classification rates of near 60% are observed. This result is a less accurate than seen when using Markov Modeling in a previous study.

1. Introduction

Currently, many accepted methods of training rely on the subjective analysis of performance by an expert. The methodology for assessing surgical skill as a subset of surgical ability is gradually shifting from subjective scoring of an expert which may be a variably biased opinion using vague criteria towards a more objective quantitative analysis. The ultimate aim is therefore to develop a modality independent methodology for objectively assessing medical competency. The methodology may be incorporated into a simulator, surgical robot, or performance tracking device during a real procedure and provide objective and unbiased assessment based on quantitative data resulting from the physical interaction between the physician and modality used to measure competency.

2. Background

Markov Modeling (MM) and Hidden Markov Modeling (HMM) are effective methods for deconstructing and understanding speech data. An analogy between spoken language and medical procedure [1] is used to apply these methods towards objective surgical skill analysis. An effective method of evaluation using MM was developed [2]. In our previous work, MM was found to successfully classify 82 subjects into two classes (expert and novice) with a 92% success rate [3]. Another study used MM to classify 5 expert and 5 novice surgeons 87.5% correctly with data acquired from the BlueDRAGON laparoscopic surgery data acquisition tool [4]. In [5], significant differences between surgeons in different levels of residency were found: 1) Magnitude of applied Forces/Torques; 2) Types of tool/tissue interactions; 3) time intervals spent in each tool/tissue interaction. Evidence was obtained which supported

Medicine Meets Virtual Reality 15, Long Beach CA, February, 2007
J.D. Weswood et. Al. (Eds.), IOS Press, 2007

the idea that a major portion of laparoscopic surgical capabilities is acquired between the first and third years of residency training.

The difference between MM and HMM is a subtle but important one. In MM, observed data is converted into model states, hence the model states directly reflect the physical reality of the process being modeled. In HMM, the states of the model do not directly reflect the physical reality. Instead, the model states represent an underlying *hidden* stochastic process that, similar to reality, could produce the observed data.

Can HMM classify subjects more correctly than MM? This study uses HMM to classify a dataset that was previously classified using MM. The results of this experiment are useful for future applications of HMM and MM to objective skill assessment algorithms.

3. Method

HMM software tools were used to construct two models from data acquired with the E-Pelvis physical simulator [6-8]. The parameters of the models were then adjusted based on obtained results to improve performance of the models. Data from 15 expert subjects and 15 novice subjects were used. These are the same data that were used as the training set in the previous MM study [2]. For more detail on HMM, the reader is referred to L.Rabiner's tutorial [9].

Parameters of expert and novice models are determined using iterative learning techniques. Subjects are scored against the two trained models by finding the probability for the most likely path through each model. The subject is classified as a member of the class for which the probability is highest.

The number of states in the model is chosen by creating 29 different models, each with a different number of states ranging from 2 to 30. The training data was then classified. The 4-state model classified the training data most correctly with the best error margin, and therefore we chose to work with this 4-state model during the rest of this experiment. The data was quantized to N clusters using K-Means clustering. Due to computational concerns, a smaller N is desirable. The distortion of the clustering process measures the MSE between the original data points and the chosen cluster centers, and this value decreases as N increases. The silhouette value is a measure of similarity between data points within a cluster relative to points in other clusters, and ranges from -1 to 1. Larger silhouette values imply a clearer distinction between clusters, and as more clusters are chosen, the silhouette value tends to decrease. N=9 was selected as a trade-off between computational concerns and distortion. For N>9, to gain more improvement in distortion would require significant loss of silhouette as compared to N<=9 (see Figure 1). The clusters correspond to observations in the HMM.

Training subject models were used to artificially generate random sample data to test the models. The experiment was performed using all subjects, and then repeated using a purified training subject subset. This subset was chosen by generating 25 trials of random sample data from individual subject models. 8 subjects' who had the highest misclassification rate were removed from the training set for the second experiment.

Medicine Meets Virtual Reality 15, Long Beach CA, February, 2007
J.D. Weswood et. Al. (Eds.), IOS Press, 2007

4. Results

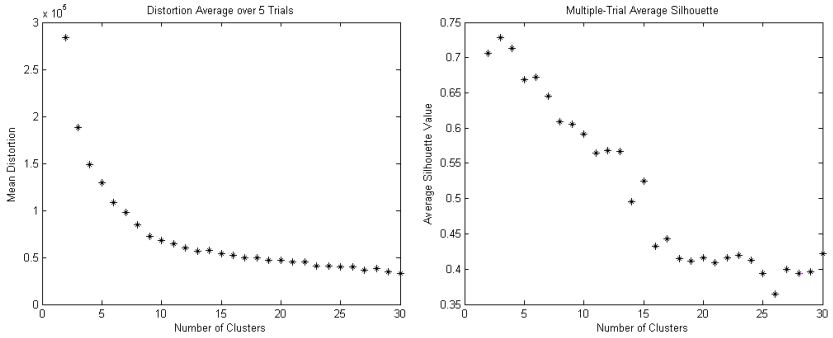

Figure 1: Plots of Distortion (Left) and Silhouette/Distortion Ratio (Right) to Select Number of Clusters

Using the unpurified training subject groups, successful classification occurred for 57% of subjects over all trials. Experts were classified correctly 87% of the time, and novices correctly 27% of the time. Using purified training subject groups, successful classification occurred for 62% of subjects over all trials. Experts were classified correctly 64% of the time, and novices correctly 59% of the time. During the selection of the number of states to use, all subjects tended to score closer to novice as the number of model states was increased. This result is not superior to the previous MM work, but HMM may have other merits.

References

[1] Rosen J., Chang L., Brown J., Hannaford B., Sinanan M., Satava R. Minimally Invasive Surgery Task Decomposition – Etymology of Endoscopic Suturing. Studies in Health Technology and Informatics, vol. 94, pp. 295-201, IOS Press, Fairfax, VA 2003.

[2] Rosen J., Brown J., Chang L., Sinanan M., Hannaford B. Generalized Approach for modeling Minimally Invasive Surgery as a Stochastic Process Using a Discrete Markov Model. IEEE Transactions on Biomedical Engineering, vol. 53, pp. 399-413, March 2006.

[3] Mackel T., Rosen J., Pugh C. Data Mining of the E-Pelvis Simulator Database: A Quest for a Generalized Algorithm Capable of Assessing Medical Skill. Studies in Health Technology and Informatics, vol. 119, pp. 355-360, IOS Press, Fairfax, VA 2003.

[4] Rosen, J., Hannaford B., Richards C., Sinanan M. Markov Modeling of Minimally Invasive Surgery Based on Tool/Tissue Interaction and Force/Torque Signatures for Evaluating Surgical Skills. IEEE Transactions on Biomedical Engineering, vol. 48, pp. 579-591, 2001.

[5] Rosen, J., Solazzo M., Hannaford B., Sinanan M. Objective Evaluation of Laparoscopic Surgical Skills Using Hidden Markov Models Based on Haptic Information and Tool/Tissue Interactions. American College of Surgeons Annual Meeting, Washington State Chapter, Lake Chelan, WA, 2000.

[6] Pugh C., Srivastava S., Shavelson R., Walker D., Cotner T., Scarloss B., et al. The effect of simulator use on learning and self-assessment: the case of Stanford University's E-Pelvis simulator. Studies in Health Technology and Informatics, vol. 81, pp 396-400, IOS Press, Fairfax, VA 2001.

[7] Pugh C., Rosen J. Qualitative and quantitative analysis of pressure sensor data acquired by the E-Pelvis simulator during simulated pelvic examinations. Studies in Health Informatics and Technology, vol. 85, pp. 376-379, IOS Press, Fairfax, VA 2001.

[8] Pugh C., Youngblood P. Free in PMC, Development and validation of assessment measures for a newly developed physical examination simulator. J Am Med Inform Assoc. 2002. 2002 Sep-Oct; 9(5):448-60.

[9] Rabiner, L. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, vol. 77, no. 2, Feb 1989.

Medicine Meets Virtual Reality 15, Long Beach CA, February, 2007
J.D. Weswood et. Al. (Eds.), IOS Press, 2007