Designing an effective intervention of motivational interview using sequence classification methods

Mehedi Hasan, BS[1], Alexander Kotov, PhD[1], April Carcone, PhD[2], Sylvie Naar-King, PhD[2]
[1]Department of Computer Science, [2]Pediatric Prevention Research Center, Wayne State University, Detroit, Michigan

Abstract *Childhood obesity is a serious public health concern in the United States and worldwide. Therefore, there is a need for identification of successful motivational interviews to facilitate the development of effective interventions for childhood obesity. This paper presents our work on the classification of utterance sequences into successful and unsuccessful behavior outcomes. Markov chain and hidden Markov models were employed and compared with a different order for the classification of utterance sequences. The classification performances achieved 75.32% and 79.89% F-measure for normal sequences and 97.36% and 96.99% F-measure for alternate sequences using 1st order Markov chain and second order hidden Markov model, respectively. This shows that the behavior codes carry some inherited information to distinguish motivational interviews. We expect that the result will improve if we use a more discriminable data set. Furthermore, the results suggest us a new way to think about designing the motivational interview.*

Introduction

In the recent times, more and more data from various domains being produced in the form of event sequences. Thus, sequence classification has become an important task in sequence mining. Sequences can contain discrete (e.g., symbolic sequences such as text, proteins, or DNA) or continuous events (e..g, time series such as ECGs or stocks), depending on the event types. Usually, sequences don't have explicit features and suffer with high dimension of feature space. Even the sequential nature of features is too difficult to capture, which make sequence classification, a more challenging task than feature-based classifications. The sequence classification methods can be divided into three large categories: feature-based classification, distance based method and model based classifications. In feature-based classification, a sequence is transformed into a single vector of features. A supervised learning algorithm such as an support vector machine[1] or decision tree[2] can be used to train the classifier. An improvement of feature-based classifier such as shapelet[3] or pattern-based[4,5] approach is used for sequence classifications. On the other hand, distance based method measure the similarity between sequences to determine the quality of the classification significance. The most commonly used distance function is euclidian distance[6], while Dynamic Time Wrapping[7] is used when more flexible matching is desired in time series data. The third category of sequence classification is model based classifiers such as hidden Markov model[8] (HMM) or other statistical models. Recently, a hierarchical approach[9] is proposed for sequence classification.

Sequence classification has a broad range of real word applications such as information retrieval, genomic analysis, health informatics, finance and abnormal detection. In genomic research, sequence classification is widely used to detect the function of a new protein[10]. In health informatics, ECG can be used as a multi-dimensional time series data to classify an individual as healthy or patient with heart disease[11]. Another example is to classify rumour stance in Twitter[12]. In information retrieval, sequence classification is employed to build models for classifying documents into different topic categories[13] such as sports, politics, news, finance and style. Sequence classification is also used in anomaly detection such as abnormal access to systems[14], malware detection[15] and money laundering[16]. Other interesting applications include classifying query log sequences to distinguish web-robots from human users[17,18]. In computer vision, sequence classification is applied for classifying images[19] and videos[20].

In this paper, our target area of application is motivational interview. In particular, this paper describes our work on the classification of utterance sequences. Classification of utterance sequences to distinguish different behavior outcomes such as positive change talk or commitment verses negative change talk or commitment, is an important part of clinical research aimed at designing effective interventions for many conditions and disorders. In this study, we focus on the transcripts of Motivational Interviews with obese adolescents (teens) and their caregivers. Childhood obesity is a serious public health concern in the United States and worldwide. Recent estimates indicate that approximately one third (31.8%) of US children age 2-19 years are overweight and 16.9% are obese[21]. Adolescents who are obese

are likely continue to be obese in adulthood and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis[22]. Therefore, there is a need for informatics-based methods to facilitate development of effective interventions for childhood obesity. One approach to designing an effective intervention is Motivational Interviewing (MI), an evidence-based counseling technique to increase intrinsic motivation and self-efficacy for health-related behavior change. The goal of a MI counseling session is to encourage patients to explore their own desires, ability, reasons, need for and commitment to the targeted behavior change. These statements, referred to as "change talk" (or CT), consistently predict actual behavior change that can be sustained for as long as 34 months after an interview.

Our previous studies[23,24] explored several machine learning methods for the automatic annotation of clinical interview fragments, which uses a codebook[25] with 115 different behavior codes. In this study, probabilistic models are employed for the analysis of sequence data to determine provider-patient communication sequences that are likely to translate into change talk and commitment language. To the best of our knowledge, there have been very limited studies on sequence classification in motivational interview such as patient-provider dialogue. A sequential analysis[26] was performed on physician-patient dialogue to find the relationship between physician and patient behaviors associated with patients' exibiting cues. Currently, a very similar work[27] was done on education data, where a sequence classification model was developed to predict the role of a student in a project team based on their communication patterns. Although hidden Markov model was originally applied to speech recognition[8], this is one of the most successful method[28–32] for sequence classification. In addition, a k-order Markov model is applied on genomic research[33] to classify protein and text sequence data. Therefore, we utilize Markov chain and hidden Markov model for the classification of utterance sequences.

Methods

Data collection

The experimental data sets for this work were constructed based on the transcripts of 37 motivational interview sessions, which include a total of 11,353 segmented and annotated utterances. Since clinical researchers tried to increase the desire and ability of adolescent to change their current behavior to target behavior, we used sequences from adolescent session and excluded sequences from caregiver session in the MI transcript. The utterances were annotated based on MYSCOPE codebook[25] with 115 different behavior codes, which are grouped into the youth, caregiver, and counselor code groups. These utterances are then scanned from first to last and divided into successful and unsuccessful sequences when a commitment or change talk is found. A total of 1072 sequences have been observed which further partitioned into two subsets based on the outcome of motivational interviews: one data set that includes all sequences of positive change talk or commitment language (880 sequences) and the other data set which includes utterances ended with negative change talk or commitment language (192 sequences). Alternate sequences are the possible combination of normal sequence where two consecutive codes can not be come from same speaker. For example, there are 6 possible alternate sequences for the following sequences: (C1, A1, A2, C2, C3, C4, A3), where adolescent and counselor codes are represented by A and C, respectively. Therefore, (C1, A1, C2, A3), (C1, A1, C3, A3), (C1, A1, C4, A3), (C1, A2, C2, A3), (C1, A2, C3, A3) and (C1, A2, C4, A3) were used as alternate sequences for the above sequence. A total of 18753 alternate sequences have been observed which further partitioned into two subsets of 18139 successful and 614 unsuccessful sequences. In our experiment, normal and alternate sequences are employed with two subsets (successful and unsuccessful) to train Markov chain and HMM. Dataset statistics and a fragment of an adolescent session transcript are presented in Table 1 and Table 2.

Table 1: General statistics of dataset.

Sequence type	Total sequences	# Successful sequences	# Unsuccessful sequences	Avg. length
Normal sequences	1072	880 (82.09%)	192 (17.91%)	5.83
Alternate sequences	18753	18139 (96.72%)	614 (3.28%)	12.89

Table 2: Fragment of the annotated transcript of a dialogue between a counselor and an adolescent.

Annotation	Description	Speaker	Text
331	Open-ended question, elicit change talk positive	Counselor	do you feel like making healthier choices for your snacks and your meals is something you would be able to do? mm-hmm meaning is that food available for you?
117	Low Uptake, positive	Adolescent	Yes
301	Structure Session	Counselor	okay and thats an important thing for us to think about cause i would not want to help you come up with a plan that you would not be able to do without somebody else help so the last part of your plan is how somebody could be supportive to you meaning how they can help you be successful and so we should choose somebody who you feel like is around often enough
112	Change Talk positive	Adolescent	my um aunt
301	Structure Session	Counselor	okay so lets stick something my aunt can do
112	Change Talk positive	Adolescent	she could when i am doing when i am eating something that i should i could not be eating but so i can choose something healthy she could tell me not to eat it
309	Affirm, low	Counselor	okay that sounds like a really great suggestion

Sequence classification techniques

Generally, a sequence is an ordered list of events. In our study, event is a behavior code represented as a symbolic value such as 117 (low uptake, positive), 331 (open-ended question), etc. Given a sequence of behavior codes represented as $S_i = \{x_1, x_2, ..., x_n\}$ and a set of class labels $L = \{l_1, l_2, ..., l_m\}$, the task of sequence classification is to learn a sequence classifier C, which is a function mapping a sequence S_i to a class label $l_i \in L$, written as, $C : S_i \to l_i, l_i \in L$. Here, L is fixed and known in advances. In our experiment, known labels are "successful" and "unsuccessful" motivational interview.

Markov Chain: In probability theory, a Markov model is a stochastic model used to model randomly changing systems where it is assumed that future states depend only on the present state and not on the sequence of events that preceded it (that is, it assumes the Markov property). Generally, this assumption enables reasoning and computation with the model that would otherwise be intractable. For the sequential analysis, we built two Markov models M and $\overline{M}$ describing provider strategies and patient responses in case of successful (M) and unsuccessful ($\overline{M}$) motivational interviews. A Markov model M can be represented as a weighted directed graph $G = (V, E, p)$, in which:

- $V = \{CML+, CHT+, CHT-, T-AMB, CCT, BLT, LUP+, LUP-, HUP-W, ...\}$ is a set of vertices, consisting of possible youth and counselor MI behavior codes;

- $E \subseteq V \times V$ is a set of edges corresponding to posssible transitions from one MI behavior code to the other in a sequence;

- $p_M : E \to [0...1]$ is a function that assigns probability $p(c_i|c_j)$ to an edge between the MI behavior codes c_i and c_j based on maximum likelihood estimator:

$$P_M(c_j|c_i) = \frac{n_{c_i,c_j}}{n_{c_i}} \tag{1}$$

where n_{c_i,c_j} and n_{c_i} are the number of times a transition between the MI behavior codes c_i and c_j and the code c_i have been observed in the training data. Given a Markov model M (such that $S \subseteq V$), a sequence of MI behavior codes $S = \{C_1, ..., C_N\}$ have been generated from a Markov model M is:

$$P_M(S) = \prod_{i=2}^{N} p_M(c_i|c_1, \ldots, c_{i-1}) = \prod_{i=2}^{N} p_M(c_i|c_{i-1}) \qquad (2)$$

Success of a given motivational interview, given a sequence of MI behavior codes corresponding to it, can be predicted based on the following formula:

$$p(S \to CT) = \log\left(\frac{P_M(S)}{P_{\overline{M}}(S)}\right) = \sum_{i=2}^{N} p_M(c_i|c_{i-1}) - \sum_{i=2}^{N} p_{\overline{M}}(c_i|c_{i-1}) \qquad (3)$$

if $p(S \to CT) > \delta$, where δ is experimentally defined threshold, the interview transcript (or a portion of it) is predicted to result in CT. To fit the model, we experimentally determined the value of δ where model provides highest F-measure with lower bias. Figure1 [this is a temporary figure, I will change it later] illustrates an example of the characteristics of Markov model by varying the value of δ.
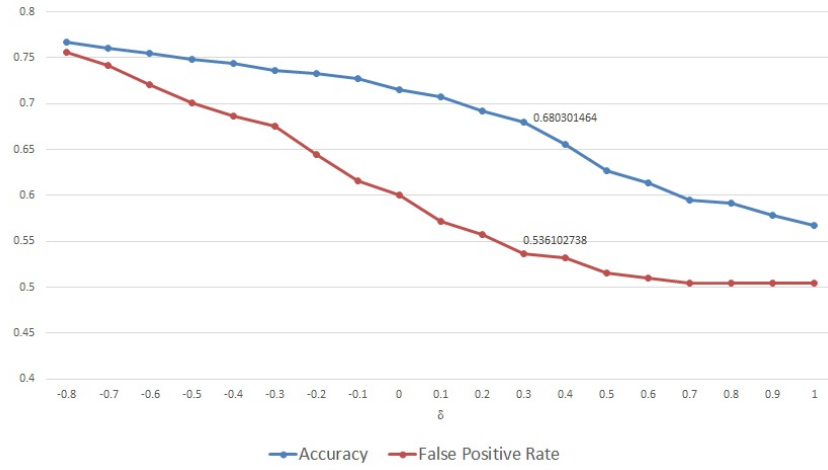

Figure 1: Characteristics of proposed Markov model by varying the value of δ

The Markov model we have discussed so far is referred to as the first order Markov model, since we only look one code in the past to compute the transition probability. A generalization of the first order Markov model is the kth order Markov chain in which the transition probability for a particular symbol x_i is computed by looking at the k preceding symbols. Thus, k^{th} order Markov chain will have N^k states each associated with a sequence of k symbols. In our experiment, we also used second order Markov model. Hence, our model have N^2 states with transitional probability matrix of size $N^2 \times N$.

Hidden Markov Model (HMM): hidden Markov models are widely used for sequence analysis because of their ability to incorporate dependencies among elements in sequence. Hidden Markov model is a very powerful tool to statistically model a process that varies in time. It can be seen as a doubly embedded stochastic process with a process that is not observable (hidden process) and can only be observed through another stochastic process (observable process) that produces the time set of observations. A HMM can be fully specified by the following quintuple:

$$\lambda = (N, M, A, B, \pi)$$

- N is the number of states for the model

- M is the number of distinct observations symbols per state, i.e. the discrete alphabet size

- A is the $N \times N$ state transition probability distribution given in the form of a matrix $A = \{a_{ij}\}$

- B is the $N \times M$ observation symbol probability distribution given in the form of a matrix $B = \{b_j(k)\}$

- π is the initial state distribution vector $\pi = \{\pi_i\}$

If we opt out the structure parameters M and N we have the more often used compact notation:

$$\lambda = (A, B, \pi)$$

This model is also called hidden for a reason. We are not required to tell it what those aforementioned states should be. We just have to tell it how many states we think are needed to model what we want. The model should be able to figure a suitable sequence of states to support our observations. The Baum-Welch algorithm was implemented to train a hidden Markov model for estimating parameters of each models using the training set. To identify the label of an unknown code sequence, the Viterbi algorithm was used to decode the sequence and compute its log probabilities respectively using HMM's trained for different classes. Then, the class with the highest log probability was assigned to the given behavior sequences. Like Markov chain, we also employed second order hidden Markov model for classification of utterance sequences.

Evaluation methods

The performance of classifiers are reported on the basis of 5-fold cross validation by using weighted macro-averaging over folds. Precision, recall, F-measure and accuracy were used to evaluate the performance of classifiers. A true positive (TP) was counted when classifier correctly classified the sequence into actual class; a false positive (FP) was counted when classifier incorrectly classified other classes to the considered class; a false negative (FN) when the considered class incorrectly classified to other classes. Precision of a class was defined as the ratio between the numbers of correctly classified sequences and all sequences identified as that particular class by the classifier. In other words, Precision = TP / (TP + FP). Recall of a class was defined as the ratio between the numbers of correctly classified sequences and all sequences of that particular class in the gold standard (labeled data set). In more specifically, Recall = TP / (TP + FN). Similarly, Accuracy of a class was also defined as the ratio between the numbers of correctly classified sequences and total number of sequences for that particular class in the gold standard. However, F-measure can be computed as the harmonic mean of precision and recall, or 2 x Precision x Recall / (Precision + Recall).

Results

Experimental evaluation of the classification of utterance sequences by using probabilistic models are reported with two set of experiments: first set of experiment was focused on normal sequences, while the second set of experiment was performed on alternate sequences.

Performance of models for the classification of normal PPC code sequences

Performance summary for the classification of normal sequences is presented in Table 3. It follows that first order hidden Markov model works very well and achieves highest precision 0.7980 with F-measure 0.7989. However, our designed second order Markov model shows lowest performance in terms of precision and F-measure. It was also noticed that the lowest recall 0.7686 achieved in first order Markov model. On the other hand, second order hidden Markov model achieves the highest recall 0.8449 and accuracy 0.8449. From Table 3, we observed that performance degrades from first order to second order. Precision and F-measures drop 6.74% & 2.39% in Markov chain and 7.27% & 2.09% in HMM, respectively, when using second order models. In contrast, recall improves 2.64%–4.84% in second order Markov model and HMM. Overall, HMM performed better than general Markov model in normal utterance sequences.

Table 3: Performance of Markov models for the classification of normal PPC code sequence. Highest performances for each metrics are highlighted in bold faces.

Model	Order	Precision	Recall	F-Measure	Accuracy
General Markov chain	First order	0.7387	0.7686	0.7532	0.7686
	Second order	0.6889	0.7889	0.7352	0.7889
Hidden Markov Model	First order	**0.7980**	0.8059	**0.7989**	0.8059
	Second order	0.7400	**0.8449**	0.7822	**0.8449**

Performance of models for the classification of alternate PPC code sequences

Table 4 shows the classification performance of models applied on alternate sequences. The results demonstrate that second order hidden Markov model performed better than Markov chain in terms of recall and achieved 97.13% recall. Second order general Markov model shows the highest precision 0.9778 as well as highest F-measure 0.9736. On the other hand, first order hidden Markov model displays lowest performance in all metrics compared to all other models. It was also examined that every models perform better in second order compared to first order. General Markov chain achieves 1%, 1% and 1% improvement for precision, recall and F-measure, respectively, from first order to second order. Similarly, hidden Markov model achieves 1%, 1% and 1% improvement for precision, recall and F-measure, respectively, from first order to second order. Overall, second order Markov model performs better for alternate utterance sequences.

Table 4: Performance of Markov models for the classification of alternate PPC code sequence. Highest performances for each metrics are highlighted in bold faces.

Model	Order	Precision	Recall	F-Measure	Accuracy
General Markov chain	First order	0.9777	0.9437	0.9604	0.9437
	Second order	**0.9778**	0.9694	**0.9736**	0.9694
Hidden Markov Model	First order	0.9392	0.9313	0.9253	0.9392
	Second order	0.9704	**0.9713**	0.9699	**0.9713**

Discussion

The purpose of this study was to facilitate the design of motivational interview by examining the performance of sequence classification in the field of health informatics. Studying the results across the different classification schemes we can make some key observations.

First, higher order models have better classification accuracy as compared to lower order models because they are able to capture longer ordering constraints present in the data set. However, the number of states in higher order Markov chains grows exponentially with the order of the chain. Consequently, it is harder to accurately estimate all the transition probabilities from the training set. That is, there are many high-order states that are not frequently observed in the training set, leading to inaccurate probability estimates. In principle, this problem can be solved if the size of the training set is increased at the same rate as the state-size of the higher order Markov chain. However, this cannot be easily done for many applications areas include our case. To skip these problems we restricted our study upto second order. Shorter average length of utterance sequence is another reason for keeping the order lower.

Second, overall performance of hidden Markov model is better than general Markov model in both type of sequences. We observed that hidden Markov model achieved near-human accuracy to categories the sequences of behavior codes. This means that hidden states capture the information passes through the codes of the utterance sequence for determining final outcome.

Third, alternate sequences provide higher performance for all Markov models. Even, it provides highest precision and F-measure for alternate sequences. This indicates that each pair of alternate behavior codes forward their mutual information to the next step more accurately which improve performance for distinguishing outcomes of the motivational interview. And this is obvious because if we know the previous two behavior codes (e.g. provider's question

and adolescent's answer) then it is easier to know the expected answer for the current question.

Finally, it was seen that successful interviews are more frequently responded by the adolescent with positive behavior changes which lead to make a positive commitment for loosing weight. Whereas, unsuccessful interviews are responded with negative behavior changes which lead to a negative commitment or no commitment.

Conclusion

It has been shown how Markov chain and hidden Markov models could be used to build classifiers for the sequence of behavior codes in motivational interviews. Experimental results shows that the behavior codes carry some inherited information to distinguish motivational interviews. This work can facilitate researchers to establish causal relationship between different communication strategies and desired behavioral outcomes without having to repeatedly wade through pages of interview transcripts. It provides information that can directly inform and increase the efficiency of clinical practice for a successful interview. This work also suggest some successful patterns for the practical use by the clinician during the interview session which will translate into positive change talk and commitment language. Direction of future work include the exploration of other methods such as BLAST[34] and Strand[35] for the classification of utterance sequences. We also have plan to use large data set in conjunction with some advanced features for the evaluation of sequence classifiers.

Acknowledgments

References

[1] Leslie C, Kuang R. Fast string kernels using inexact matching for protein sequences. Journal of Machine Learning Research. 2004;5(Nov):1435–1455.

[2] Chuzhanova NA, Jones AJ, Margetts S. Feature selection for genetic sequence classification. Bioinformatics. 1998;14(2):139–143.

[3] Ye L, Keogh E. Time series shapelets: a new primitive for data mining. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2009. p. 947–956.

[4] Kudenko D, Hirsh H. Feature generation for sequence categorization. In: AAAI/IAAI; 1998. p. 733–738.

[5] Lesh N, Zaki MJ, Ogihara M. Mining features for sequence classification. In: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 1999. p. 342–346.

[6] Keogh E, Kasetty S. On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and knowledge discovery. 2003;7(4):349–371.

[7] Keogh EJ, Pazzani MJ. Scaling up dynamic time warping for datamining applications. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2000. p. 285–289.

[8] Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989;77(2):257–286.

[9] Nallam K, et al. An Effective Implementation of Hierarchical Approach For Sequence Classification. IJACTA. 2016;3(2):143–146.

[10] Deshpande M, Karypis G. Evaluation of techniques for classifying biological sequences. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer; 2002. p. 417–431.

[11] Wei L, Keogh E. Semi-supervised time series classification. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2006. p. 748–753.

[12] Lukasik M, Srijith P, Vu D, Bontcheva K, Zubiaga A, Cohn T. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In: Proceedings of 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2016. p. 393–398.

[13] Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR). 2002;34(1):1–47.

[14] Lane T, Brodley CE. Temporal sequence learning and data reduction for anomaly detection. ACM Transactions on Information and System Security (TISSEC). 1999;2(3):295–331.

[15] Drew J, Hahsler M, Moore T. Polymorphic malware detection using sequence classification methods and ensembles. EURASIP Journal on Information Security. 2017;2017(1):2.

[16] Liu X, Zhang P, Zeng D. Sequence matching for suspicious activity detection in anti-money laundering. In: International Conference on Intelligence and Security Informatics. Springer; 2008. p. 50–61.

[17] Duskin O, Feitelson DG. Distinguishing humans from robots in web search logs: preliminary results using query rates and intervals. In: Proceedings of the 2009 workshop on Web Search Click Data. ACM; 2009. p. 15–19.

[18] Tan PN, Kumar V. Discovery of web robot sessions based on their navigational patterns. In: Intelligent Technologies for Information Analysis. Springer; 2004. p. 193–222.

[19] Li J, Najmi A, Gray RM. Image classification by a two-dimensional hidden Markov model. IEEE transactions on signal processing. 2000;48(2):517–533.

[20] Zhang J, Xu C, Jing P, Zhang C, Su Y. A Tensor-Driven Temporal Correlation Model for Video Sequence Classification. IEEE Signal Processing Letters. 2016;23(9):1246–1249.

[21] Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of obesity and trends in body mass index among US children and adolescents, 1999-2010. Jama. 2012;307(5):483–490.

[22] General US. Surgeon Generals vision for a healthy and fit nation. Washington, DC: HHS. 2010;.

[23] Kotov A, Hasan M, Carcone A, Dong M, Naar-King S, BroganHartlieb K. Interpretable probabilistic latent variable models for automatic annotation of clinical text. In: AMIA Annual Symposium Proceedings. vol. 2015. American Medical Informatics Association; 2015. p. 785.

[24] Hasan M, Kotov A, Carcone AI, Dong M, Naar S, Hartlieb KB. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. Journal of biomedical informatics. 2016;62:21–31.

[25] Carcone AI, Naar-King S, Brogan K, Albrecht T, Barton E, Foster T, et al. Provider communication behaviors that predict motivation to change in black adolescents with obesity. Journal of developmental and behavioral pediatrics: JDBP. 2013;34(8):599.

[26] Eide H, Quera V, Graugaard P, Finset A. Physician–patient dialogue surrounding patients expression of concern: applying sequence analysis to RIAS. Social Science & Medicine. 2004;59(1):145–155.

[27] Jaber M, Wood PT, Papapetrou P, González-Marcos A. A Multi-Granularity Pattern-Based Sequence Classification Framework for Educational Data. In: Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on. IEEE; 2016. p. 370–378.

[28] Mutsam N, Pernkopf F. Maximum margin hidden Markov models for sequence classification. Pattern Recognition Letters. 2016;77:14–20.

[29] Eickeler S, Kosmala A, Rigoll G. Hidden markov model based continuous online gesture recognition. In: Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on. vol. 2. IEEE; 1998. p. 1206–1208.

[30] Srivastava PK, Desai DK, Nandi S, Lynn AM. HMM-ModE–Improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. BMC bioinformatics. 2007;8(1):104.

[31] Won KJ, Prügel-Bennett A, Krogh A. Training HMM structure with genetic algorithm for biological sequence analysis. Bioinformatics. 2004;20(18):3613–3619.

[32] Chai W, Vercoe B. Folk music classification using hidden Markov models. In: Proceedings of International Conference on Artificial Intelligence. vol. 6. sn; 2001. .

[33] Yakhnenko O, Silvescu A, Honavar V. Discriminatively trained markov model for sequence classification. In: Data Mining, Fifth IEEE International Conference on. IEEE; 2005. p. 8–pp.

[34] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215(3):403–410.

[35] Drew J, Hahsler M. Strand: fast sequence comparison using mapreduce and locality sensitive hashing. In: Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. ACM; 2014. p. 506–513.