CSE497-Engineering Project I

PROJECT SPECIFICATION DOCUMENT

IDENTIFYING NON-STATIONARY SEQUENCES BY VARIABLE ORDER MARKOV CHAINS

STUDENT: Feyza Nur TOPÇU-150110075

ADVISOR: Borahan TÜMER, ASSOC.PROF.

.....

CO-ADVISOR: Peter SCHÜLLER, ASST.PROF.

.....

1. Problem Statement

In my project, I employ variable order Markov chains (VOMCs) to analyze non-stationary and noisy sequences and attempt to identify the repeating cycle and/or significant structures related to each individual subsequence. Here I assume that the non-stationary sequence under analysis is composed of a series of noisy but stationary subsequences. To start with, I implement a fast and adaptive version of VOMCs as presented in [1]. Then I employ this VOMC to analyze the non-stationary sequence composed of tokens from a finite alphabet. The final VOMC is represented in the form of a probabilistic suffix tree (PST) where each subtree of its root is another VOMC related to one subsequence. In that PST, to find the change points in non-stationary sequences which are used by VOMCs, we want to use statistical information based on stochastic learning-based weak estimation as presented in [7].

2. Problem Description

VOMCs which are studied at a diversity of application domains are an improved class of well-known Markov Chains (MCs).

According to [5], *"MCs are a mathematical system that undergoes transitions from one state to another on a state space."* MCs are modeled by graphs where vertices and their usually weighted interconnections represent the states from the state space under analysis and their stochastic transitions, in respective order, and the next state in this model can be predicted by looking at the relevant context (i.e., the series of incoming data of some length). The context length determines the order of MCs; if the probability of the next state depends only upon the current state then we work on first order MCs (FOMCs), in general if it depends on the first most recent i states prior to the next state then we work with i^{th} order MCs. If we study on the previous tokens based on the context of a specific length of two, then we use second order MCs (SOMCs). The FOMCs and SOMCs are used on wind speed modelling [6], to analyze web navigation for users [5] etc. However, we cannot find the difference between two subsequences in the original sequence if these two subsequences' conditional probabilities are the same for all combinations. For example, assume that the sequences S_1 and S_2 are *abac* and *abacacab*, respectively.

The first order probabilities for S_1;

P(a)	½
P(b)	¼
P(c)	¼

$P(a\|a)$	0
$P(b\|a)$	½
$P(c\|a)$	½
$P(a\|b)$	1
$P(b\|b)$	0
$P(c\|b)$	0
$P(a\|c)$	1
$P(b\|c)$	0
$P(c\|c)$	0

Table 1

The first order probabilities for S_2 ;

$P(a)$	½
$P(b)$	¼
$P(c)$	¼
$P(a\|a)$	0
$P(b\|a)$	½
$P(c\|a)$	½
$P(a\|b)$	1
$P(b\|b)$	0
$P(c\|b)$	0
$P(a\|c)$	1
$P(b\|c)$	0
$P(c\|c)$	0

Table 2

As shown in the example, these two different sequences cannot be distinguished using a first order MC. For problems where we are required to study the context of variable lengths VOMCs are employed as a suitable variant of MCs.

As mentioned before, the subsequences have repeating cycles. If these subsequences are identical, then there will no problem to identify the original sequence. However, if we study real world sequences then we should be aware that we deal with noisy data which implies that our solution should be robust to noise. For example, the ECG records can be non-stationary sequences for a problem that is relevant with VOMCs [9].

VOMCs can be represented by probabilistic suffix trees (PSTs). These are sometimes called Context Trees, since the sequences are the "context" of the analysis. To produce PSTs, we use there is a known method that is the AB algorithm developed by Apostolico and Bejerano. This algorithm is also the building block for my project [2].

VOMCs are used in a diversity of application areas such as machine learning, information theory, physics, chemistry [10], testing, coding and data compression, speech recognition, queueing theory [11], internet applications [12], economics and finance [13], social sciences [14], mathematical biology [15], genetics, games, music [16], baseball [17], Markov text generators and others [5].

VOMCs have several advantages over other methods such as HMMs: they are more explanatory and they can still perform well even in case of scarce training sets. The drawback of the VOMCs is the higher cost, which increases depending on the increasing of the order level. Hidden Markov Models (HMMs) have relatively lower level of complexity; their disadvantages are they need domain understanding and require large training sets. Despite their higher cost, VOMCs are commonly used on various application areas.

The reason for my choosing this project is that I am specifically interested in the research areas of general machine learning and in particular sequential decision making. MCs and VOMCs are popular mathematical tools used in sequential decision making. Further they have a variety of application areas as mentioned above. I want to create an implementation of VOMCs, in order to produce a solution for this kind of problems. Using the outcome of this project we can for example predict the creator of a song (musical signal analysis and classification/recognition) or the writer of some text (text processing and classification), we can learn and predict protein families (protein classification) or heart diseases according to patient records (biological/biomedical signal analysis).

I am planning to make a literature survey about VOMCs to prepare myself for the implementation of VOMCs and the application of VOMCs to music processing and text classification to produce novel results in these domains.

3. Aims of the Project

- Conduct research on VOMCs: I am reading some articles about VOMC and I am planning to acquire comprehensive knowledge on this area. Moreover I want to read some related works which are implemented for some problems and relevant with VOMCs in real life.

- Implementation of VOMCs: This is my major aim in my project. During this process, I am planning to implement a variant of VOMCs and I want this implementation to yield correct results within reasonable execution times.
- Applying VOMCs to analysis of music:
    - Combining with music chords: In this part of the project, I am planning to use input files which contain music chords in form of vectors. Given this data, I am planning to use VOMCs to predict the creator of the songs with acceptable accuracy.
    - Combining with Turkish words: In this aim of the project, I am planning also use Turkish words as input and use VOMCs to predict suffixes and roots for those words.

4. Related Work

The most relevant work is about how to learn and classify proteins in linear time and space, whichwas developed by Alberto Apostolico and Gill Bejerano [2]. They introduced the AB algorithm which is a known and accepted method to realize VOMCs. This method creates an automaton for the given set of sequences and can observe the changes in the sequences using this automaton. After that, they can predict the next token of the sequence using collected probability information.

The aim of creating the automaton is to reduce the execution time to linear time. For example for a given training set *S* of sequence requires $\Theta(Ln^2)$ worst case time where *n* is the total length of sequence, *L* is the length of a longest substring of *S* to be considered for a candidate state in the automaton.With a learning automaton, for any *L*, this will take *O(n)* time [2].

In the AB algorithm, the tree-shaped variants of probabilistic automata called Probabilistic Suffix Trees (PSTs) are used. They present automata equivalent to PSTs with some properties such as linear learning time. The conditional empirical probability is a mathematical concept that defines the probability that some token in the given context can occur given the occurrence of some token(s). The PSTs are created with regard of these conditional empirical probabilities. In this algorithm the learning time should be *O(n)*, since there are some methods by which enhanced PSTs are created within linear time. Those methods are the *similarity pruning* and *support pruning*. By these methods, the PST will be pruned from some non-conditional nodes and we will obtain an enhanced PST and the learning algorithm will be in linear time. I plan to use this algorithm to implement VOMCs.

Another work about VOMCs is Bayesian classifiers for genomic signatures developed by Daniel Dalevi [3]. The genomic signatures have alignments of homologous sequences.

The VOMCs also use an input file which contains similar sequences, so this area is appropriate for using VOMCs. They reject a false hypothesis of horizontal gene transfer by proposing an alternative method. VOMCs are used to determine probabilities of the next transition, using the alphabet of nucleotides (A,T,G and C). This prediction can be made by creating a PST for this alphabet. This PST contains roots and leaves where the leaves point to next states. Thus, we can capture the dependencies of the DNA chromosomes and genomic signatures by PSTs. They also proposed a novel method for this topic and we can observe differences between two methods.

Another works on protein classification area for VOMCs is the development of a method for the problem of protein domain detection. This work is also developed by Gill Bejerano [4]. In this work, the input tokens originating from an alphabet of nucleotides of DNAs are unaligned groups of protein sequences. The method is developed and it chops those sequences up in form of groups which share the same underlying statistics. The PST is created using those segments.

To create an enhanced PST, some calculations and prunings are made on PST. With a enhanced PST, we can obtain that regions of similar statistics by matching a unique to each domain. For the classification of proteins the most comman statistical approach is the Hidden Markov Models (HMMs). The alternative models to HMMs is the probabilistic finite automata, the Variable Memory Markov (VMM). They have several advantages over HMMs. One of them is the VMM captures longer correlations and higher order statistics of the sequence. The second is VMMs use PST, VMMs can be learned by the automaton with an optimal sense. As mentioned in AB algorithm, VMMs are efficient about learning in linear time. Since, the protein classification is made using VMMs as in this work. They developed the PST for the clusters of the proteins using VMMs and the output of the algorithm has specialized in recognizing a certain protein region.

5. Scope of the Project

My first goal for this project is implementation of VOMCs. VOMCs classify non-stationary subsequences which might contain noise. For testing purposes, we can produce analogous subsequences with tokens stemming from a finite alphabet. The production of such non-stationary subsequences with noise is not in the scope of this project. I will use this part from outside, provided by my advisor.

VOMCs are mathematical systems which symbolize the tokens taken from real world domain. In the literature there exist mathematical formulas for calculating the probabilities,

to create PSTs and to prune the PSTs for simplification. I will exploit these probabilities and conditional probabilities for generating the PSTs' information and I further use the pruning methods which can generate enhanced PSTs. In this project, I will have to use the mathematical formulae specifying the pruning methods as proposed in [1]. In this part, I am using those equations from outside.

During my project, I need some other input files like some probabilities about inputs like co-variances for subsequences, also the probabilities lie in the basic of co-variance calculations. I am also using those input files which are made by my advisor.

In contrast to those, the other calculations and implementations I will make by myself. First, I will try to implement VOMCs for one non-stationary sequence and analogous to this subsequence with a noise. I will try to create a PST for those subsequences and I will prune the PST for the non-conditional nodes until I end up with the final version of the enhanced PST. By this way, we can find the original sequence. As the next step for my project, I will try to implement VOMC for several non-stationary subsequences which are different forms and ranges. I will try to find the change points that show us the point where a sequence ends and another starts. After I find unify those PSTs those points, I will create enhanced PSTs for each subsequence and I will with a super root. So, I have several PSTs for each subsequence and I can find the original sequence.

After the implementation of VOMCs, I will apply them to music accords and Turkish words. Those files are also used from outside, I will not create them myself during the project. Therefore I will apply my project to a real world application and hopefully find some new results in these problem domains.

6. Success Factors and Benefits

In my project, the noise factor is an parameter to obtain the program is succesful or not. The relation with noise factor and my project's success is that; whereby noise we can produce different subsequences from original sequence. We use input files include sequences with different noises. For each input file we calculate the succesful rate that is number of defined sequences over total number of sequences.For example, we assume there are 1000 sequences and our program defined 850 sequences with %5 noise. We can say our program is successful with %85 (850/100) for %5 noise.

Another indicator is time for my project. The time is important because, the program should notice the changes between subsequences rapidly. If my project provides this ability, I can say it recognizes the changes quickly and with minimum execution time.We should obtain the sequences with converge time.

If the results of this project are correct and fast, we can apply it to real-world problems. I and my advisor are planning to combine this project with music accords which should be in form of vector to use for this project. I will not implementation this part of project just I will use those vectors to modify with my project to give like an input file.

Moreover, I, my adrvisor, and my co-advisor will try to apply the software created in this project to an existing collection of Turkish words. The noise in this data comes from irregularities of natural language. Here our goal is to create PSTs for each word and to try to identify roots of words and grammatical suffixes of words without human help.

7.Methodology and Technical Approach

In my project, I will use articles about VOMC models which provide mathematical equations for the implementation of VOMCs in several application areas [1]. Those equations will be useful for my project implementation. The major algorithm will applied for this project is the AB algorithm [1,2].

During my project implementation, I will obtain real-world data input files which contain probabilities of tokens, co-variances, Turkish words, and music accords in form of vectors from my advisor and my co-advisor.

The C language will be used during implementation of the project. I prefer the Linux operating system for developing with the C language, because of that I will use Fedora like working area.

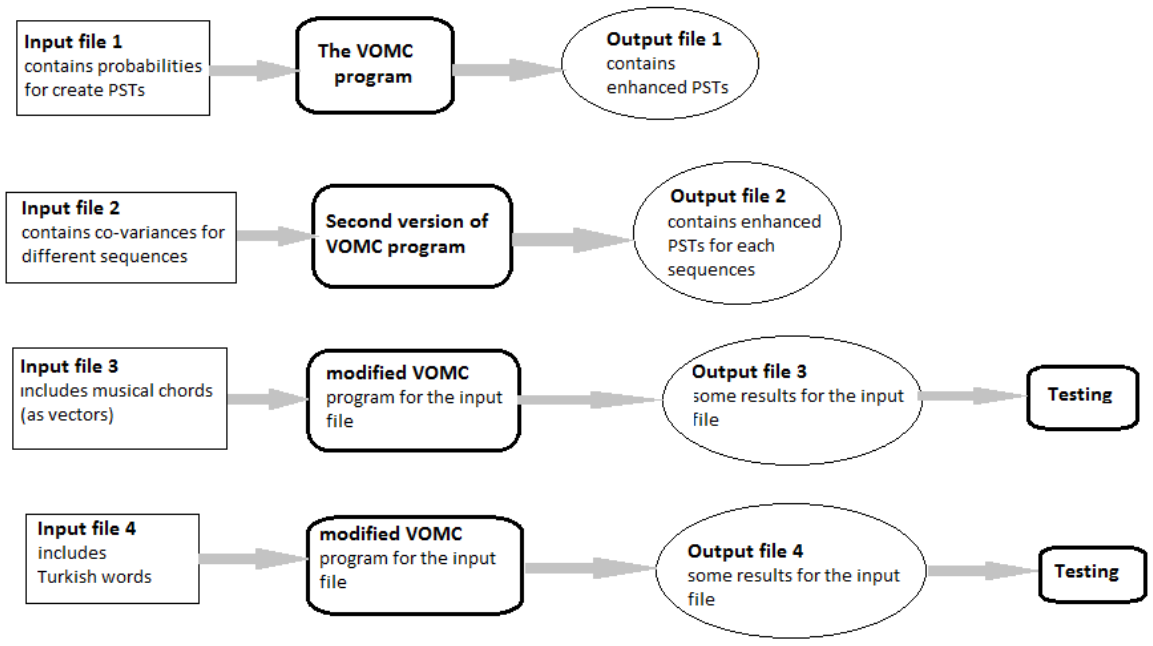
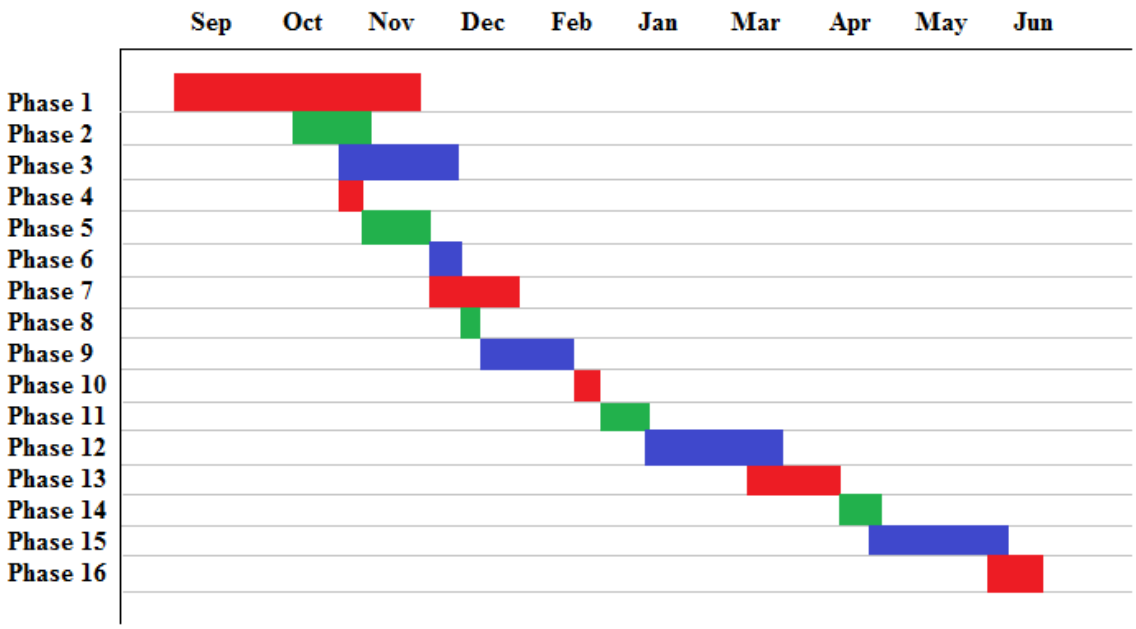


Figure 1

8.Management Plan

There will be several phases during the project. These will be in outline as follows:

- **Phase 1:** Literature Survey: In this phase, the articles which are related with VOMCs and their application domains are read to understand the VOMCs.
- **Phase 2:** Implementation of some basic programs: after the literature survey to better understand of the VOMCs, I will writing some codes which make easy to implementation VOMCs.
- **Phase 3:** Project Specification Document: there will be preparing of this document with my advisor.
- **Phase 4:** Getting some input documents for my project contains probabilities for 1st version implementation of VOMC.
- **Phase 5:** I will begin to implementation of VOMC.
- **Phase 6:** This phase will be testing phase for the measure the success of implementation.
- **Phase 7:** There will be preparing for the representations such as create slides and documents for this progress. There will be also Progress Report in this phase.
- **Phase 8:** Getting input documents contains co-variances and sequences are generated from noise by my advisor.
- **Phase 9:** In this phase, I will try to implementation of the second version of the VOMC which is the suitable for the subsequences which are formed different ranges and lengths.
- **Phase 10:** There will be testing of the implementation.
- **Phase 11:** Getting input file contains music accords which are in form of vectors.
- **Phase 12:** I will modify the implementation for the suitable to input file.
- **Phase 13:** Testing phase of implementation.
- **Phase 14:** Getting input file which contains Turkish words from my co-advisor.
- **Phase 15:** I will modify the implementation for the input file which contains Turkish words helping with co-advisor.
- **Phase 16:** Testing phase the final version of implementation with my advisor and co-advisor.



9. References

[1]. MH. Schulz, D. Weese, T. Rausch, A. Döring : Fast and adaptive variable order Markov chain construction . Algorithms in Bioinformatics. 8(306-317)(2008)

[2]. Apostolico, A., Bejerano, G.: Optimal amnesic probabilistic automata or how to learn and classify proteins in linear time and space. J. Comput. Biol. 7(3-4) (2000)

[3]. D. Dalevi, D. Dubhashi, M. Hermansson: Bayesian classifiers for detecting HGT using fixed and variable order markov models of genomic signatures. Bioinformatics. Oxford Univ Press .22 (5): (517-522) (2006)

[4]. G. Bejerano, Y. Seldin, H. Margalit, N. Tishby: Markovian domain fingerprinting: statistical segmentation of protein sequences . Bioinformatics . Oxford Univ Press. 17 (10): (927-934) (2001)

[5]. http://en.wikipedia.org/wiki/Markov_chain

[6]. A Shamshad, MA Bawadi, WMA Wan Hussin, TA Majid: First and second order Markov chain models for synthetic generation of wind speed time series. Energy. Elsevier. 30(693-708)(2005)

[7]. BJ Oommen, L Rueda, "Stochastic learning-based weak estimation of multinomial random variables and its applications to pattern recognition in non-stationary environments," Pattern Recognition .39(328-341) (2006)

[8]. Z Yuan: Prediction of protein subcellular locations using Markov chain models
. FEBS letters.Elsevier.451 (23-26) (1999)

[9]. L Senhadji, L Thoraval, G Carrault: Continuous wavelet transform: ECG recognition based on phase and modulus representations and hidden Markov models. Wavelets in medicine and …, - books.google.com. (439-440) (1996)

[10]. Kutchukian, Peter; Lou, David; Shakhnovich, Eugene (2009). "FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules occupying Druglike Chemical". Journal of Chemical Information and Modeling 49 (7): 1630–1642.

[11]. S. P. Meyn: Control Techniques for Complex Networks, Cambridge University Press, (2007).

[12] Page, Lawrence and Brin, Sergey and Motwani, Rajeev and Winograd, Terry: The PageRank Citation Ranking: Bringing Order to the Web (Technical report). (1999).

[13] Hamilton, James: A new approach to the economic analysis of nonstationary time series and the business cycle. Econometrica (Econometrica, Vol. 57, No. 2) 57(2): 357–84. (1989).

[14] Acemoglu, Daron; Georgy Egorov; Konstantin Sonin: Political model of social evolution. Proceedings of the National Academy of Sciences 108: 21292–21296. (2011).

[15] George, Dileep; Hawkins, Jeff Friston, Karl J., ed. Towards a Mathematical Theory of Cortical Micro-circuits. PLoS Comput Biol 5 (10). (2009).

[16] K McAlpine, E Miranda, S Hoggar: Making Music with Algorithms: A Case-Study System. Computer Music Journal 23 (2): 19. (1999).

[17] B Bukiet, ER Harold, JL Palacios: A Markov chain approach to baseball. Operations Research(1997) 45.(14-23)