Work in progress: mining exceptional sequences using log likelihood based quality measures

Rianne Margaretha Schouten, Wouter Duivesteijn, and Mykola Pechenizkiy

Eindhoven University of Technology, Eindhoven, the Netherlands
{r.m.schouten, w.duivesteijn, m.pechenizkiy}@tue.nl

Abstract. Modeling sequential data with a first order Markov chain gives valuable insights about the average starting and transition behaviour of sequences. In heterogeneous data, it is likely that some sequences follow alternative or exceptional patterns. We aim to find those sequences using log likelihood based quality measures in Exceptional Model Mining. We empirically show that parameter based measures do not perform well with an increasing number of states. In contrast, several log likelihood based measures detect both exceptional starting behaviour and exceptional transition behaviour. In addition, these measures are robust with respect to the number of states. We conclude by introducing open questions.

Keywords: Exceptional sequential behaviour · First order Markov chains · Log likelihood based quality measures

1 Introduction

Markov models in all their variants are frequently used to mine patterns in sequential data [1, 5, 7]. A first order Markov chain in particular gives valuable insights about the average behaviour of sequences of discrete values or states [3, 6]. In such a Markov model, a vector of initial probabilities models the starting state of a sequence and a transition matrix models the transition between any source and target state.

However, it is often likely that some heterogeneity in the data exists and that some sequences have alternative starting or transition behaviour. We use the framework of Exceptional Model Mining (EMM) [2] to explore this heterogeneity and to discover exceptionally behaving sequences. In particular, we evaluate the quality measures as proposed by [4]. We find that those measures do not perform well with an increasing number of states and can discover only exceptional transition behaviour. Our experiments show that several log likelihood based quality measures, on the other hand, are able to detect exceptional starting behaviour too. In addition, these measures are robust for the number of states.

In the following, we discuss our choices for the quality measures and outline the design of the experiment. After discussing our findings, we introduce open questions.

2 Approach

We consider sequences of transitions between state values $v_1, v_2, ..., v_S$. For N independent and identically distributed sequences of length T, a 1^{st} order Markov chain models the joint probability distribution as follows:

$$p(X) = p(\mathbf{x}_1, ..., \mathbf{x}_t, ..., \mathbf{x}_T) = p(\mathbf{x}_1) \prod_{t=2}^{T} p(\mathbf{x}_t | \mathbf{x}_{t-1}). \tag{1}$$

Here, for $t \in \{1, ..., T\}$, $\mathbf{x}_t$ is a column vector of length N containing the states at time t. The probability of starting in state v_s, $p(\mathbf{x}_1 = v_s)$, is modeled with initial probabilities vector $\pi = [\pi_1, ..., \pi_S]$ where $\sum_{s=1}^{S} \pi_s = 1$. The transitions are modeled with an $S \times S$ transition matrix $\mathbf{A}$ where cell $a_{ij} = p(\mathbf{x}_t = v_j | \mathbf{x}_{t-1} = v_i)$ and $\sum_{j=1}^{S} a_{ij} = 1$ for all $i, j \in \{1, ..., S\}$. The model is called memoryless, as the target state at time $t \in \{2, ..., T\}$ depends only on the source state at time $t-1$.

We aim to identify sequences in the dataset with any form of unusual sequential behaviour. We therefore use an instance of Exceptional Model Mining (EMM) [2]: a local pattern mining method seeking subgroups of the dataset behaving somehow exceptionally. Here, this exceptional behaviour is measured in terms of parameters of a first order Markov chain with time dependent attribute X as the target concept. EMM employs a rule based description language: resulting subgroups are described as a conjunction of attribute-value pairs. For instance, in our experiment the description of the true subgroup is '$z_0 = 1 \wedge z_1 = 1$'.

Note that when a time dependent target attribute is used, and the descriptive attributes are time-dependent as well, subgroups will be defined on the transition level and only parts of a sequence will end up in the candidate subgroup. Because we intend to find subgroups of entire sequences, our descriptive attributes are on the sequence level.

2.1 Related work

An EMM model class for first order Markov chains exists [4]. This model class focuses on finding subgroups of transitions, rather than subgroups of sequences. To assess exceptionality, the Manhattan distance between the transition matrix of a candidate subgroup (A^{SG}) and the data transition matrix (A^D) is employed:

$$\delta_{tv} = \sum_{i=1}^{S} \sum_{j=1}^{S} \left| a_{ij}^{SG} - a_{ij}^{D} \right|. \tag{2}$$

An adapted version in [4] skews the search process away from tiny subgroups with spurious exceptionalities:

$$\omega_{tv} = \sum_{i=1}^{S} \left(w_i \sum_{j=1}^{S} \left| a_{ij}^{SG} - a_{ij}^{D} \right| \right). \tag{3}$$

Here, w_i is the number of transitions in the subgroup with source state v_i.

2.2 Log likelihood based quality measures

If sequences differ from the overall data regarding their starting state, quality measures such as δ_{tv} and ω_{tv} (Equations (2), (3)) are unlikely to find them; these measures focus only on the transition matrix. In this work, we show that log likelihood based quality measures are able to detect exceptional sequences, also when exceptionality stems from initial probabilities.

Log likelihood based quality measures have been used [8] in the context of Hidden Markov Models (HMMs), for instance by applying an evaluation based quality measure called *weighted divergence*:

$$\varphi_{\mathrm{WD}}(SG) = \sum_{n \in SG} \left(\log f_{SG}(x_n) - \log f_D(x_n)\right). \tag{4}$$

Here, f_{SG} indicates the likelihood of a model that is trained on the subgroup and applied to every sequence x_n in the subgroup. This measure is evaluation based, because it compares the likelihood of a subgroup between a local (f_{SG}) and global (f_D) model. Another version called *approximate Kullback-Leibler divergence* [8] divides by the subgroup size:

$$\varphi_{\mathrm{KL}}(SG) = \frac{\varphi_{\mathrm{WD}}}{|SG|} = \frac{1}{|SG|} \sum_{n \in SG} \left(\log f_{SG}(x_n) - \log f_D(x_n)\right). \tag{5}$$

Semi-evaluation based quality measures compare the likelihood of the subgroup under the global model with the likelihood of the entire data under the global model. This is referred to as *relative log likelihood* [8]. This quality measure is positive when, on average, the subgroup has a better fit than the entire data; since we are also interested in subgroups with a lower fit, we will use the absolute relative log likelihood, and a weighted version:

$$\varphi_{\mathrm{AbsRL}}(SG) = \left| \frac{\sum_{n \in SG} \log f_D(x_n)}{|SG|} - \frac{\sum_{n=1}^{N} \log f_D(x_n)}{N} \right| \tag{6}$$

$$\varphi_{\mathrm{WAbsRL}}(SG) = |SG| \cdot \varphi_{\mathrm{AbsRL}}. \tag{7}$$

In the context of Dynamic Bayesian Networks (DBNs), BIC has been proposed [1, p. 130] as the basis of a *mismatch score*. The authors replace f in Equation (4) with the BIC and compare the subgroup to its complement. We compare the subgroup to the entire dataset:

$$\varphi_{\mathrm{BIC}}(SG) = BIC_{SG}(SG) - BIC_D(SG). \tag{8}$$

Here, $BIC_{SG}(SG)$ is the BIC of a model trained on (subscript) and applied to (brackets) the subgroup. Note that the normal BIC multiplies the log likelihood with -2. Because the beam search algorithm optimizes for larger values of the quality measure, we multiply with 2 instead. Note as well that the number of parameters in a first order Markov model is $K = S^2 - 1$.

3 Experimental results

We compare the effectiveness of several (types of) quality measures in mining exceptional sequences in a dataset: the parameter-based quality measures in Equations (2) and (3) as introduced by [4], the evaluation based measures from Equations (4) and (5), the semi-evaluation based measures from Equations (6) and (7) (originally introduced by [8], adapted for the task at hand), and the evaluation based quality measure from Equation (8) (which is an adaptation of the mismatch score from [1]).

The synthetic datasets are generated with $N \in \{100, 1000\}$ sequences, $S \in \{2, 5, 25\}$ states and $T \in \{2, 5, 25\}$ time points. An $S \times S$ transition matrix is randomly drawn from a uniform distribution. Each row $i \in \{1, ..., S\}$ is scaled such that $\sum_{j=1}^{S} a_{ij} = 1$. The initial probability vector, π, is also drawn from a uniform probability distribution and scaled to sum to 1. We generate sequences by first sampling the starting state with probabilities π and then sampling subsequent states with the probabilities of the transition matrix.

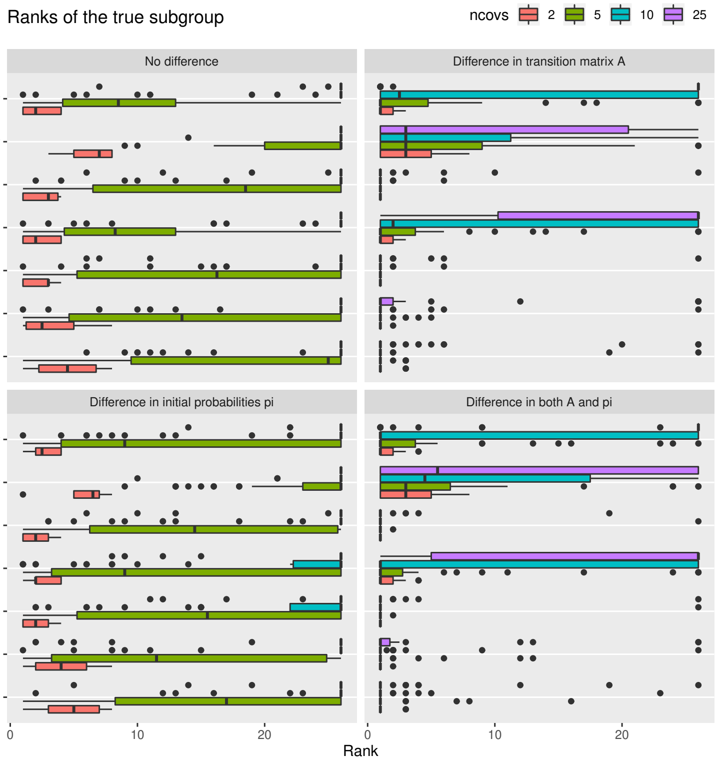

Fig. 1. Boxplots (50 repetitions) of ranks of the ground truth subgroup for parameter setting $N = 100$, $T = 25$ and $S = 25$. The panels show various types of exceptionality.

Furthermore, we sample $ncovs \in \{2, 5, 10, 25\}$ binary descriptive attributes Z. As mentioned earlier, each descriptive attribute is a sequence level attribute. For all $z \in Z$, $p(z_n = 1) = 0.5$ for all $n \in \{1, ..., N\}$. Furthermore, we sample $nreps = 50$ datasets for every parameter combination.

In every dataset, a ground truth subgroup SG is created for sequences with description '$z_0 = 1 \wedge z_1 = 1$'. We create four types of exceptionality: 1) another transition matrix is randomly sampled, and subgroup sequences are drawn using that matrix, 2) another initial probability vector is randomly sampled, and subgroup sequences are drawn using that vector, 3) both the transition matrix and the initial probabilities differ and 4) there is no difference.

Next, we perform a beam search [2, Algorithm 1] with $w = 10$, $d = 5$, $q = 25$, constraining the subgroup size to $\geq 0.1 \cdot N$ (see [2] for more information about these choices) and apply the quality measures as defined above. For each run, we store the rank of the true subgroup as returned by the algorithm. When the true subgroup is not part of the top-25, we assign a rank of 26.

Figure 1 shows the ranks of the ground truth subgroup as boxplots for parameter setting $N = 100$, $T = 25$ and $S = 25$. The four panels show the results for the various types of exceptionality as defined above. Figure 2 presents two of these panels for parameters $N = 100$, $T = 5$ and $S = 5$.

The experimental results confirm our intuition that log likelihood based quality measures detect subgroups of sequences for all investigated types of exceptionality. Subgroups with an exceptional pattern of initial probabilities cannot be found with parameter based quality measures like δ_{tv} and ω_{tv}, since these measures focus only on differences between transition matrices. Especially evaluation based measures φ_{WD} and φ_{BIC} detect exceptionalities in initial probabilities well.

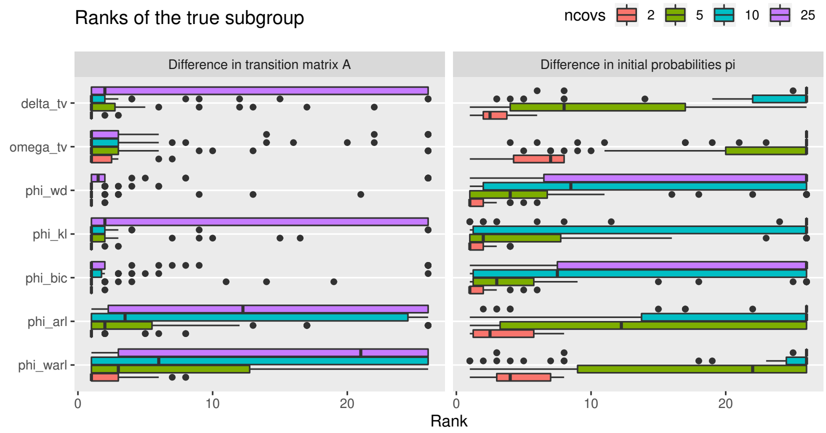

Fig. 2. Boxplots (50 repetitions) of ranks of the ground truth subgroup for parameter setting $N = 100$, $T = 5$ and $S = 5$. Left: exceptionality in terms of the transition matrix. Right: exceptionality in terms of initial probabilities.

In addition, we find that δ_{tv} and ω_{tv} do not perform well with an increasing number of states (compare the top right panel in Figure 1 with the left panel in Figure 2). This effect increases when the number of covariates increases. In contrast, most of the likelihood based quality measures rank the true subgroup first despite of the small number of sequences and the large number of states and covariates. In particular, the boxplots of φ_{WD} and φ_{BIC} entirely collapse on the first rank and have barely any outliers (Figures 1 and 2).

4 Conclusions and discussion

Our results from the controlled experiments on synthetic datasets show that exceptional sequential behaviour, modeled by a first order Markov chain, can be detected by applying Exceptional Model Mining with sequence level descriptive attributes and evaluation based quality measures. In contrast to parameter based quality measures such as δ_{tv} and ω_{tv}, quality measures φ_{WD} and φ_{BIC} are robust and detect exceptionality in both starting behaviour and transition behaviour.

The chance of finding sequences that are exceptional in their initial probabilities decreases with an increasing number of transitions, even when the transition behaviour itself is not exceptional. Hence, the open question is: how can we detect exceptionality in initial probabilities for longer sequences?

The reason that semi-evaluation based quality measures such as φ_{AbsRL} and $\varphi_{\mathrm{WAbsRL}}$ performed badly is possibly because they use a global model that suffers from noise. Hence, an interesting question is: in what situations are semi-evaluation based quality measures still valuable for detecting exceptionality?

Weighted versions of other quality measures perform worse than their origins. We would like to understand the effect of weighting more thoroughly.

References

1. Bueno, M.L.P.: Unraveling Temporal Processes using Probabilistic Graphical Models. Ph.D. thesis, Leiden University (2020)
2. Duivesteijn, W., Feelders, A.J., Knobbe, A.: Exceptional model mining. Data Mining and Knowledge Discovery **30**(1), 47–98 (2016)
3. Kiseleva, J., Thanh Lam, H., Pechenizkiy, M., Calders, T.: Discovering temporal hidden contexts in web sessions for user trail prediction. In: WWW'13. pp. 1067–1074 (2013)
4. Lemmerich, F., Becker, M., Singer, P., Helic, D., Hotho, A., Strohmaier, M.: Mining subgroups with exceptional transition behavior. In: KDD'16. pp. 965–974 (2016)
5. Meier, J., Dietz, A., Boehm, A., Neumuth, T.: Predicting treatment process steps from events. Journal of Biomedical Informatics **53**, 308–319 (2015)
6. Sadagopan, N., Li, J.: Characterizing typical and atypical user sessions in clickstreams. In: WWW'08. pp. 885–894 (2008)
7. Singer, P., Helic, D., Taraghi, B., Strohmaier, M.: Detecting memory and structure in human navigation patterns using markov chain models of varying order. PloS one **9**(7), e102070 (2014)
8. Song, H.: Model-Based Subgroup Discovery. Ph.D. thesis, Univ. of Bristol (2017)