Unsupervised detection of interesting discrete sequences using Bayesian surprise

Ken McGarry (ken.mcgarry@sunderland.ac.uk)
University of Sunderland https://orcid.org/0000-0002-9329-9835
Sharon McDonald
University of Sunderland
Dominic Button
University of Sunderland

Research Article

Keywords: Bayesian surprise, Probabilistic Suffix Tree, Anomaly, interestingness, Pattern

Posted Date: May 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1238817/v1

Cognitive Computation manuscript No.
(will be inserted by the editor)

Unsupervised detection of interesting discrete sequences using Bayesian surprise

Ken McGarry · **Sharon McDonald** · **Dominic Button**

Received: date / Accepted: date

Abstract *Background / introduction*
In this work we employ Bayesian surprise to detect interesting/anomalous patterns from discrete sequence data. Many domains consist of discrete sequential time-series such as DNA analysis, online transactions, web click-stream navigation, cyber-attacks, financial transactions and especially sociology life-course data. The difficulty is that each data set has its own unique characteristics and many anomalies defy categorization. Since anomalies are by nature infrequent and elusive, we often do not have enough data for a supervised approach. However, novelty and surprise play a fundamental role in human and animal behavior for survival, attention and adaptation.

Methods
We use three sequence datasets (Swiss Health, Sepsis, and BioFamilies) which are are composed of simpler motifs which are used to build Probabilistic Suffix Trees (PST) which can capture complex relationships based on motif location and frequency of occurrence. New data that deviates from established motifs either in location of appearance, frequency of appearance, or motif composition may represent recurring patterns that may be different in some way. Bayesian surprise is the result of mismatches between our expectations and actual results, hence the degree of surprise or anomalousness attached to a pattern will vary with respect to these differences.

Results
Each data set is assessed by Bayesian Surprise and several other criteria, providing indications of why certain patterns are interesting and why others are not.

Conclusions
Bayesian surprise can detect data with other properties that would be missed by information theoretic measures such as Shannon surprise and entropy for example.

Keywords Bayesian surprise · Probabilistic Suffix Tree · Anomaly · interestingness · Pattern

Introduction

The ability to be surprised is fundamental to many human cognitive and intellectual endeavors, it is an essential trait for learning and discovering new knowledge [4, 5]. There is general agreement between cognitive scientists that surprise is an emotion that arises when differences occur between expectations and actual results [10, 15]. This mismatch can be accounted for in a principled way using Bayes theory, which is perfectly suited to update beliefs in the light of new information. We implement a modification of Bayesian surprise discussed by Itti which corresponds to subjective beliefs that are re-

*Ken McGarry, School of Computer Science, University of Sunderland, UK, ken.mcgarry@sunderland.ac.uk

vised as new information appears [26]. Bayes theorem allows the conversion of our prior beliefs into posterior beliefs. Therefore, Bayesian surprise is a criterion to judge discrepancies between a systems prior and posterior beliefs on any given matter. The bigger the discrepancy then the more surprised we should be [27]. The use of surprise and novelty is also finding applications in goal directed learning when developing automated systems [45, 21] and agent based systems [42, 35, 2]. Furthermore, product design has benefited from this approach whereby *surprise* is used as a creative metric to predict if customers will find new features and styles exciting and attractive [7].

Bayesian surprise naturally lends itself to anomaly detection which seeks to uncover patterns or trends in data that differ from normal expectations. Anomalies can be either noise or potentially interesting and useful patterns. However, because of their rarity they may simply be dismissed as noise. Furthermore, anomalies cannot always be expected to follow previously known patterns or trends. In this work we tackle the difficulty of detecting the trend and predicting anomalies ahead of time. The process of detecting outliers/anomalies (interesting patterns) attempts to discriminate normal from abnormal behaviors or patterns and often requires access to domain knowledge. The activity is further complicated when noise and anomalous patterns are similar, so filtering cannot always be used to remove noise since this approach may screen out interesting anomalies.

In this work we use data sets consisting of sequences of symbols that will form recurring patterns or motifs that imply important changes or activities in the time-series. For example, a shop may collect data on customer activities, each possible activity will be identified by a letter as shown in table 1. We can see that transactions 1-2 are normal in that customers enter the shop and pick items and either pay or replace items back on shelf and then leave the shop but transaction 3 is fraudulent, it is different, not expected and therefore surprising.

Table 1: Example strings

Transaction	String
1.	A, B, B, B, C, D
2.	A, B, B, E; C, D
3.	A, B, B, D

Where: A=customer entered shop; B=picked up an item; C=purchased items; D=left shop; E=put item back

We use three discrete symbol based data sets which are split into test/train partitions. Probabilistic suffix trees (PST) are trained, and then provide probabilities on the test data. The differences between the probabilities of what the PST has learned and the expected outputs is used as the prior and posterior relationships. Differences between them are measured by the Bayesian Surprise criteria to determine unique or anomalous patterns.

Related work

The SEQUITUR system of Nevill-Manning builds a structure from sequences of discrete symbols by replacing repeated phrases with a grammatical rule that generates the phrase, and continuing this process recursively. SEQUITUR operates by abstracting subsequences that occur more than once into rules and continuing this operation recursively, the result is a hierarchical rule based system [39].

Previous work by Lin and Keogh in detecting motifs in time-series led to the development the Piecewise Aggregate Approximation (PAA) algorithm and the Symbolic Aggregate approXimation (SAX) algorithm [34]. PAA is a very popular algorithm used to convert the time series into discrete levels or symbols. The sequences/subsequences of symbols may form patterns or motifs that represent particular activity in the time series data [28]. Further improvements on PAA and SAX symbolic representation method for discretization includes preserving the information contained in the slope characteristics of the time series segments [49]. Other Computer Science areas such as temporal association rules [48] have similar issues such as modelling the temporal aspects by integrating interval-based relationships to occurrences of items in the database.

Natural Language Processing and speech recognition provide many insights into modelling sequence information, especially when dealing with strings, letter and word occurrences [43, 47]. For example part-of-speech-tagging (POS) allows the sequence of words to be represented to resolve ambiguity. Hidden Markov Models (HMM) are able to label text data in POS applications, we describe HMM in detail later. DNA searching algorithms have elements in common with motif detection, they all search for recurring or interesting patterns in sequential, discrete data [32]. A common data structure to hold sequential data is the suffix tree. Huang *et al* used suffix trees to contain temporal data for periodic patterns of different types (full, inner, and tail patterns) [25]. Work by Cohen used segmentation by information theory to decompose strings [13]. Experiments conducted by Reick on the suitability of various data structures such as sorted arrays, tries and suffix trees for sequence analysis provided valuable insights into their usage [43].

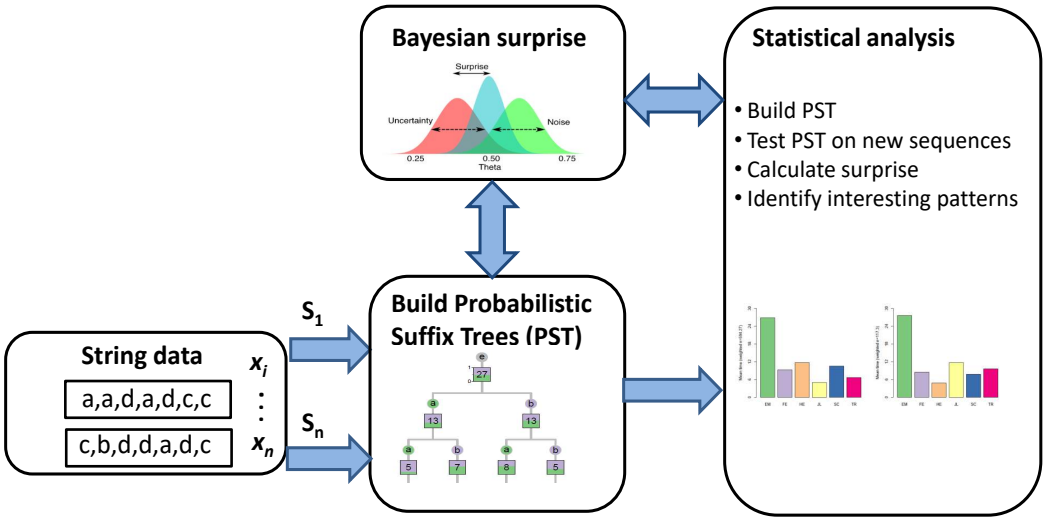

Fig. 1: Bayesian surprise and Probabilistic Suffix Tree representation

Hidden Markov Models (HMM) are often used to capture the dynamics of sequence data [41]. Furthermore, a Markov chain is a model we can employ to inform us about the probabilities of sequences of variables, events or states, each of which can take on values from an alphabet. The alphabet can be words, or tags, or symbols representing the problem domain. A Markov chain makes a very strong assumption that if we want to predict the future in the sequence, all that matters is the current state [11]. All the states before the current state have no impact on the future except via the current state. The HMM is considered to be the model of choice for sequential problems. [33] and is used in sequence anomaly detection [18]. In this work we use a variable length Markov chain (VLMC) implemented as a Probabilistic Suffix Tree (PST) to hold the sequences.

Probabilistic Suffix Trees

Probabilistic Suffix Trees, these are widely used in sequence modeling [1, 30, 8]. The Markov approaches are used in several studies to solve anomaly detection problems with the idea that an odd behavior might be represented not only by a single observation, but also by a series of consecutive observations [50]. The Probabilistic Suffix Tree (PST) is a representation of the Variable Length Markov Chain (VLMC) and uses a suffix tree as its storage structure. The VLMC property enables differing lengths of sequences to be used when training and testing the PST, rather than to be constrained by fixed length sequences which is a major disadvantage.

The key property of Markov Chains is that they are *memory-less*, i.e., each state depends only on the previous state. So we can immediately define a probability for the next state, given the current state:

$$P(x_i|x_{i-1}, ..., x_1) = P(x_i|x_{i-1})$$

Therefore, the columns of A have to sum up to 1. In this way, the probability of the sequence can be decomposed into:

$$P(x) = P(x_L, x_{L-1}, ..., x_1) = P(x_L|x_{L-1})P(x_{L-1}|x_{L-2})...P(x_2|x_1)P(x_1) \quad (1)$$

$P(x_1)$ can also be calculated from the transition probabilities: If we multiply the initial state probabilities at time $t = 0$ by the transition matrix, we get the probabilities of states at time $t = 1$ and therefore we also have them for time $t = n$.

VLMCs are able to model complex sequential data and do not require complex estimation procedures to learn from data. They have superior performance and applications to both MCMs and the more common Hidden Markov Models (HMMs) [12]. However, one interesting feature is the predictive or generative ability to compute a probability distribution for the learned sequences. Given the model S built on the training data we can calculate the likelihood of a new test sequence given a model S is called the sequence prediction. The test sequences are passed into the PST which will return the conditional probabilities for the expected next symbols in the prediction. We rely on equation 2 below to generate these short sequences, those sequences with low

probabilities may be anomalies or at least interesting to the user in some way.

$$\forall \ell \in \{0, 1, 2, ...\} : \sum_{x \in A^{\ell}} P^{S}(x) = 1. \quad (2)$$

where: A is the alphabet of symbols and S is the generative model (built on the alphabet) representing the probability A^{ℓ}. While $x \in A^{\ell}$ is the sequence presented to the PST (S). These probabilities are used by the next stage to determine if a pattern or sequence is surprising or interesting.

Bayesian surprise and novelty

The connection between human reasoning and Bayesian modeling is the assumption that Bayesian cognitive theories are effectively a rational analysis grounded on the observations of an initial theory and revising it based on new data [31]. There is sufficient evidence for assuming the Bayesian approach for making models of cognition is essentially correct even though there are several counter arguments where humans deviate from Bayesian inference [23, 31]. The evidence is based on several human problem solving tasks that produce consistent results when tasks become too difficult to manage using normative techniques and thus become reliant on heuristic approaches [3]. Numerous attempts to devise algorithms to detect interesting patterns using surprise and novelty as criteria [37]

The usual convention for stating Bayes rule is given below:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (3)$$

Where:

- $P(h|D)$ is the posterior probability of the hypothesis h given the data D
- $P(D|h)$ is the likelihood of D given h
- $P(h)$ is the prior probability of hypothesis h
- $P(D)$ is the marginal likelihood of the probability of the data D

We implement the Bayesian surprise measure S, which tests the two-fold variation between prior and posterior over the hypothesis and data and returns a value [4, 27]. This value will be either positive or negative depending on the observers belief in the hypothesis when it either increases or decreases. The *distance* measure is the Kullback-Liebler divergence measure. Several applications have recently used the Bayesian Surprise criteria to as part of a feedback criteria for improving the reliability of machine learning models such as neural networks and thematic maps [24, 14, 22].

$$S(D, h) = distance[P(h), P(h|D)] \quad (4)$$

The Bayesian surprise measure provides a natural and useful method for defining and representing novel and surprising patterns [5]. Equation 5 calculates the distribution over all hypothesis $h \in \mathcal{H}$. The surprise is given as the two-fold difference between $P(h|D)$ and $P(h)$.

$$\begin{aligned} S(D, \mathcal{H}) &= distance[P(h), P(h|D)] \\ &= \sum_{\mathcal{H}} P(M|D) log \frac{P(M|D}{P|M} \end{aligned} \quad (5)$$

Novelty and surprise play a fundamental role in human and animal behavior for survival, attention and adaptation. Surprise is not however entirely related to the information content of a pattern alone [6, 27]. Experiments with patterns of visual white-noise (random but with high information content) presented to participants over time, their Bayesian surprise quickly decreased and soon vanished as they adjusted their beliefs so that the random patterns are anticipated and expected. "*Thus, more informative data may not always be more important, interesting, worthy of attention, or surprising*" [4]. Shannon or similar information theoretic measures would erroneously classify the majority of unusual patterns as surprising because of their low probability.

The Kullback-Liebler (KL) divergence examines the relative entropy [29] and is the most suitable approach for measuring the difference between prior and posterior distributions [9]. It is defined by:

$$KL[p(y)||p(x)] = \sum_{i=1}^{n} p(y_i) log \frac{p(y_i)}{p(x_i)} \quad (6)$$

Where: $p(y)$ represents the posterior or correct distribution of data and $p(x)$ represents the hypothesis or model. We obtain a value measuring the difference between the prior distribution $p(\theta)$ to the posterior distribution $p(\theta|y)$ [46]. The machine learning perspective of a novel pattern is deemed to be a statistical outlier that is different to the probability density function of previously observed patterns [36], in other words novel patterns are those with low estimated probability of occurrence.

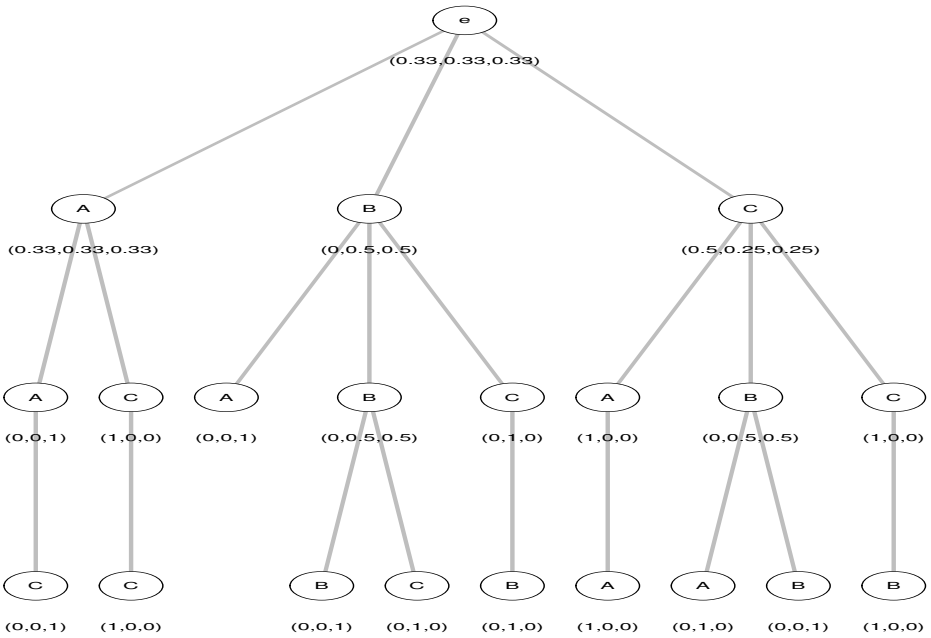

Fig. 2: Example Probabilistic Suffix Tree based on the training sequence "A-B-C-B-B-B-C-C-A-A-C-A". The tree is of maximal depth three with a minimal value of one occurrence of a symbol for it to be included. The values are the probabilities of the symbols occurrence.

However, without some sort of memory or the ability to recognize previous interesting patterns, each presentation of data would result in similar outcomes of interesting scores being assigned. This can be tackled in one of two ways, either incorporate the new data patterns into the PST and retrain it or use a decay function over time that will dampen surprise such as that proposed by Baldi [4].

$$\frac{1}{a_n} + log(1 - \frac{1}{a_N + b_N}) \approx \frac{1-p}{p_N} \tag{7}$$

Where: N is the number of data samples, a and b refer to the respectively to the prior and posterior probability values.

Data

We now define the notations for representing the data which comprise the alphabets, strings, letters and motifs. The data sets consist of a series of categorical data sequences of fixed length and also variable length vectors. The sequences represent a series of actions or situations that have occurred. The datasets also contain variables such as age, gender, biological measurements using a variety of data types. However, we do not use these and concentrate on the discrete event sequence data. In table 2 a summary of the datasets is given. We remove records with missing values from the Swiss Health and BioFam data.

Table 2: Summary of the three data sets

Dataset	Records	Sequence length	Symbols
Sepsis	1,000	Variable(3-30)	16
Swiss Health	2,612	Fixed(11)	5
BioFam	2,000	Fixed(16)	8

Biofam data set. The Biofam (Bio-Family) data consists of 2000 individuals monitored over 16 years and represents family life sequences from the retrospective survey carried out by the Swiss Household Panel (SHP) in 2002. The data set contains sequences of family life states from age 15 to 30 (sequence length is 16) and a series of covariates. The covariate data is not used in our experiments. The states numbered from 0 to 7 are

defined from the combination of five basic states, namely Living with parents (Parent), Left home (Left), Married (Marr), Having Children (Child), Divorced:

- Parent
- Left
- Married
- Left+Marr
- Child
- Left+Child
- Left+Marr+Child
- Divorced

In table 3 an example of six, 11 year sequences is given.

8cm

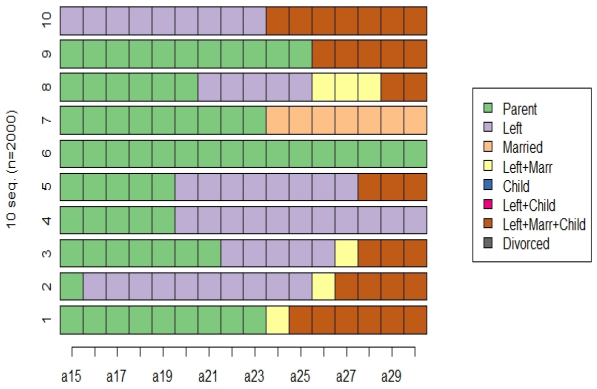

Fig. 3: Sequence ordering of biofam symbols

Sepsis data set This data set consists of events of sepsis cases from a hospital. Sepsis is a life threatening condition usually caused by an infection. One case represents the pathway through the hospital. The events were recorded by the ERP (Enterprise Resource Planning) system of the hospital. There are 1,000 individual patient records with in total 15,000 events that were recorded for 16 different activities. A further, 39 data attributes are recorded, such as the results from several tests conducted by the Doctors and the particular lab where the tests were made, along with any medication given. The main feature of interest of this dataset is the variable length of the discrete sequence records.

- • AdmissionC • AdmissionNC • CRP
- • ERRegistration • ERSepisis Triage • ERTriage
- • IVAntibiotics • IVLiquid • LacticAcid
- • Leucocytes • ReleaseA • ReleaseB
- • ReleaseC • ReleaseD • ReleaseE
- • ReturnER

In table 4 an example of the variable length sequences is shown, the first column is the ID for that particular patient. We only show the first eight sequences, the average sequences length is 13, the shortest 3 and the longest is 33. The patient journey always starts with ERReg and terminates with one of the five Release types (typically ReleaseA), with the possibility of returning to ER (Emergency Room) should the infection return.

8cm

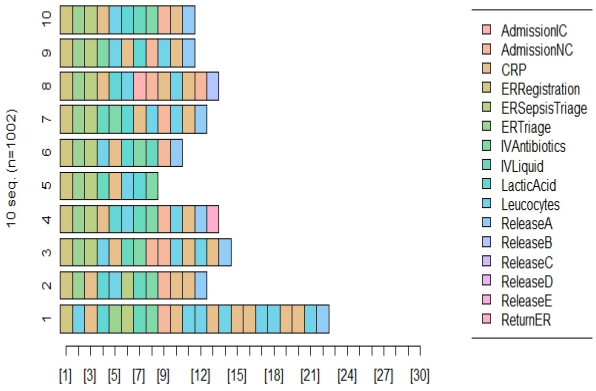

Fig. 4: Sequence ordering of sepsis symbols

Swiss Health Data The self-rated health (SRH) data set contains sequences for 2,612 respondents of a survey conducted by the Swiss Household Panel (SHP). The individuals were aged between 20 and 80 years at the start of the survey. The data is organized into 11 variables of 2,612 records (one reading for each person over the 11 years of the survey 1999-2009). The survey has missing data. Respondent's self rated health is collected at each yearly wave of the SHP with the following question: "How do you feel right now?" Possible answers are: very well; well; so, so (average), not very well and not well at all.

- G1 (very well)
- G2 (well)
- M (so, so (average))
- B2 (not very well)
- B1 (not well at all)
- * (missing)

Table 3: Example of six biofamily records for 16 consecutive years, the first column refers to the recorsd id number, where: (P) indicates living with Parents, (L)Left home, (LM) Left home and Married, (LMC) Left Home and Married with Children,

ID	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2012	2013	2014	2015	2016
1335	P	P	P	P	P	P	P	P	P	L	LM	LM	LMC	LMC	LMC	LMC
1516	P	P	L	L	L	L	L	L	L	L	L	L	L	L	L	L
1870	P	P	P	P	P	P	P	P	P	LM	LM	LM	LM	LM	LM	LM
2162	P	P	P	P	P	P	P	P	P	P	P	LM	LM	LM	LM	LMC
398	P	P	P	P	L	L	L	L	L	LM	LMC	LMC	LMC	LMC	LMC	LMC
902	P	P	P	P	P	P	LMC	LMC	LMC	LMC	LMC	LMC	LMC	LMC	LMC	LMC

Table 4: Example of six Sepsis data records

ID	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
841	ERReg	ERTriage	LacticAcid	Leucocytes	CRP	ERSepsisTriage	IVLiquid	IVAntibiotics
825	ERReg	ERTriage	ERSepsisTriage	IVLiquid	Leucocytes	CRP	LacticAcid	IVAntibiotics
430	ERReg	ERTriage	CRP	Leucocytes	LacticAcid	ERSepsisTriage	IVLiquid	IVAntibiotics
95	ERReg	ERTriage	ERSepsisTriage	CRP	Leucocytes	IVLiquid	IVAntibiotics	AdmissionNC
209	ERReg	ERSepsisTriage	ERTriage	IVLiquid	IVAntibiotics	CRP	Leucocytes	LacticAcid
442	ERReg	ERTriage	ERSepsisTriage	CRP	Leucocytes	LacticAcid	IVLiquid	IVAntibiotics

In table 5 an example of six records is shown, the first column is the unique ID for that persons 11 year record. The next columns are the responses to the questionnaire over the 11 year period.s

Fig. 5: Sequence ordering of Swiss health symbols

Methods

For all data sets we remove records with missing data and divide the training/test split randomly by 75/25. Our approach is not classification, it is unsupervised in the sense that test data should be viewed as as new patterns encountered by our system which needs to determine how different these patterns are compared with the trained tree. The Probabilistic Suffix Trees (PST) are constructed on the training sequences and provide an overall probability score for each symbol in the data set, we use the PST package and the TraMineR package [20, 19] for creating the sequence format required by the PST. The test data is passed through the PST and for each symbol in every test record the difference between the probabilities is noted. The 1st record is the probabilities assigned to the training data symbols as deemed by the PST. The 2nd record is the probabilities for the test data as they are passed through the PST.

In algorithm 1 the Input takes a data structure consisting of either fixed length or variable length strings . The algorithm will Output a trained PST and the probabilities for each symbol in the training data. The Parameters are used to train the PST and do not require much in the way of tuning. The L parameter is an integer value and sets the maximal depth of the PST. The *nmin* parameter is an integer value and controls the minimum number of occurrences of a string to add it in the tree. The parameter *ymin* is also an integer value and controls the smoothing for conditional probabilities, assuring that no symbol, and hence no sequence, is predicted to have a null probability. The parameter *ymin* sets a lower bound for a symbol's probability.

In lines 1-2, the string data is converted into special sequence format (for the PST) and then split 75/25 into train and test partitions. In line 3, the PST is training on this data using the parameters. Lines 4-5 generate the probabilities for each unique symbol in the PST, the *prob* parameter can be set to probabilities or relative frequen-

Table 5: Example of six Swiss Health Data records for 11 years, where: (G1) indicates very well, (G2) well, (M) average health, (*) missing data

	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
1491	G1	G2	G2	G2	G2	G1	G2	G2	G2	G2	G2
2202	G2	*	G1	G2	G2	G2	G2	G2	G2	G2	G2
1503	G2	M	M	M	M	G2	M	M	M	M	G2
1411	G1	*	G1	G1	G2	G1	G2	G1	G1	G1	G2
322	G2	M	G2	G2	G2	G1	G2	G2	G2	G2	G2
1299	M	G2	G2	M	G2	G1	G2	G2	G2	G2	G2

cies. Line 6 returns the trained PST and the associated symbol probabilities.

Algorithm 1 Train Probabilistic Suffix Tree

Input: set of string data D
Output: Trained PST $\mathbf{PST}_n$; Probabilities for each symbol $\mathbf{P_{sym}}$
Parameters: $[L = 10; nmin = 2, ymin = 0.001]$
1: Convert $Seq \leftarrow D$ using seqdef().
2: Split $Seq_{train}, Seq_{test} \leftarrow Seq$ by 75/25.
3: Train PST using $pstree([Parameters])$ on Seq_{train}
4: Obtain probabilities for each $\mathbf{P_{sym}} \leftarrow$ PST
5: cprob(PST , L = 0, prob = TRUE)
6: **Return** $[\mathbf{P_{sym}}; \mathbf{PST_n}]$

In algorithm 2 the trained PST, equations 5 and 7 use the test data to generate probabilities. For input the algorithm receives a trained PST, probabilities for each symbol in the PST and the test data. It will output upon completion for each test record: the decay values β_n; the Bayesian Surprise $Surp_n$; Entropy E_n; KullBack-Leibler distance KL_n. In lines 1-2, the values for KL_n, E_n and $Surp_n$ are set to zero, these will be calculated for each and every record in the test data. The Prior and Posterior values are set to zero, the likelihood value is set to a binomial function that is centered at 0.5. This is our main assumption (belief) in our approach, it assumes that overall, any given pattern has a 0.5 probability of being surprising. Where $P(Data|\theta)$ is the likelihood, $P(\theta|Data)$ is the prior, and $P(\theta)$ is the marginal likelihood.

Lines 4-6, perform the Bayesian inference stage by calculating the likelihood, the Posterior and the standardised Posterior. The posterior must be standardised, this is an important property of any probability density or mass function is that it integrates to one. The likelihood refers to the probability of observing the data that has been observed assuming that the data came from a specific scenario. The posterior can be computed from three key values: 1. A likelihood distribution, P(Data—θ) 2. A prior distribution, P(θ) 3. The average likelihood.

Lines 7-10 calculate the Bayesian surprise $Surp_n$; the Shannon Entropy E_n; KullBack-Leibler distance KL_n and decay values β_n. These values are stored in vectors for later use for comparisons. Using decay, how many patterns are genuinely interesting prior to reaching the zero value cut-off? This is a process based on similar patterns reappearing over time, as they are presented. Line 11, reiterates the loop until all of the test patterns have been analysed. Line 12 returns the results which are passed to algorithm 2.

Algorithm 2 Calculate Bayesian Surprise

Input: Trained PST $\mathbf{PST}_n$; Probabilities for each symbol $\mathbf{P_{sym}}$; Test data $\mathbf{T}_n$
Output: $\beta_\mathbf{n}$; $\mathbf{TP_n}$; Bayesian Surprise $\mathbf{Surp_n}$; Entropy $\mathbf{E_n}$; KullBack-Leibler $\mathbf{KL_n}$
1: Initialize $\mathbf{KL_n}; \mathbf{E_n}; \mathbf{Surp_n} = 0$
2: Initialize $P(Data|\theta) = 0$, $P(\theta|Data)$=0, $P(\theta)$=dbinom(0.5)
3: **repeat**
4: Obtain the Likelihood $P(Data|\theta) = P_{sym_n}$.
5: Calculate the Posterior $P(\theta|Data) = P(Data|\theta)$ x $P(\theta|Data)$
6: Calculate the standard Prior $P(\theta) = P(Data|\theta)$ x $P(\theta)$
7: Calculate $Surp_n$ = mean(Posterior) - mean(Prior) **Eqn 5**
8: Calculate $KL_n = KL(Prior, Posterior, log_{10})$
9: Calculate $E_n = shannon.cond.ent(Prior, Posterior, log_{10})$
10: Calculate $\beta_n = i$ **Eqn 7**
11: **until** $P_{sym} \notin T_n$
12: **Return** $[\mathbf{Surp_n}; \mathbf{KL_n}; \mathbf{E_n}; \mathbf{TP_n}, \beta_\mathbf{n}]$

After calculating the entropy, KL distance and Bayesian Surprise for the three datasets we determine if a test pattern is surprising or not. However, it is probably more important to determine why a given sequence is surprising. This is based use a number of measures to analyze the structure, composition and regularity of the interesting sequences. We have six methods that assess the sequences:

1. Sequence entropy. Shannon entropy is used to measure the diversity of the states or symbols in any sequence, based on the length of sequence and the number of symbols in the alphabet [40]. A more varied sequence will have a higher entropy than a sequence composed of fewer symbols.
2. Sequence complexity. A sequence may be defined in terms of the complexity of distinct sub-sequences that can be discovered from the distinct state sequences and is often called turbulence in the literature[16].
3. Longest Common Prefix (LCP). Distance measures from string theory are used to compute similarities and distances. The longest common prefix for the se-

quences is the common prefix between the two most dissimilar strings.[17].

4. Longest Common Sub-sequence (LCS). A sub-sequence is an relaxation of the idea of a sub-string, a sub-sequence is a pattern that appears in the same relative order, but is not necessarily contiguous [44]. This method is particularly well suited to DNA symbol matching.
5. Optimal Matching Distance (OMD) and Clustering. In string theory, edit distances are generated by optimal matching that seeks the minimal cost, in terms of insertions, deletions and substitutions. This is mainly a process of transforming one sequence into another and is widely used in bioinformatics using the Needleman-Wunsch algorithm. [38]. The computations produce a matrix that can be clustered for further information.
6. Sequence event transitions. Rather than simply examine the sequences of symbols, we can also observe the sequences of transitions or events between sequences deemed as surprising and those that are not. The actual transitions between symbols might reveal why they are of interest.

We make our source code (written in the statistical language R) and data available from:
https://github.com/kenmcgarry/BayesSurprise

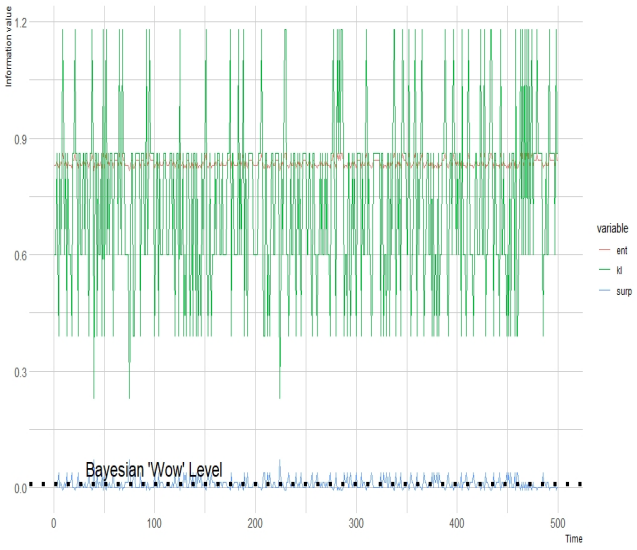

Fig. 6: KL, Entropy and Bayesian surprise on Biofam Data without interest decay

Results and Discussion

We now compare the values of entropy, Kullback-Liebler (KL) distance and Bayesian surprise for each test data set. All three measures are based on the differences between prior and posterior probabilities. It should be noted that the line plots have a more of less regular appearance. This is because similar patterns occur in the test data, the smaller the number of unique symbols, the more likely the symbol sequences will be similar and hence repeated over time. The "Wow" level is the two-fold difference of the Bayesian Surprise value when derived from the differences between the prior and posterior values as devised by Itti [27]. This is calculated for each dataset and will be unique in each case. It should be noted that the Bayesian Surprise is a much more conservation measure, especially when coupled with the "wow" threshold which requires a two-fold increase of Bayesian Surprise for any pattern in the test data to be considered interesting. Entropy, in all three datasets has a much higher response to the differences between prior and posterior values, i.e. the overall information gain is high. The KL distance is more conservative than entropy but still can be extremely variable across the three datasets.

Viewing the Biofam data results, shown in fig 6, the information value of the KL measure fluctuates between a range of 0.3-1.2 and thus has high information content. Similarly entropy has high information content but does not fluctuate to any great extent. The Bayesian Surprise measure for Biofam is just over the zero value, the "Wow" level is also just above zero. Shown in Fig **??**, the regularity of the Biofam data is evident, 10 records are shown each with 16 sequences. The majority of the individuals (younger people) in the study are still living with parents but in many sequences we can observe life changes over the period of 16 years as the individuals marry, have children, move from home and parents etc.

Examining the Sepsis data results, shown in fig 7, the information value of the KL measure fluctuates between a range of 0.0-0.5. However, in this data entropy has a much higher information content than the other measures and fluctuates between 0.75-1.00. The Bayesian Surprise measure for Sepsis is between 0.1-0.35 and has similar characteristics to the KL measure. The "Wow" level is around 0.23. Shown in Fig 4, the varied size of the sequences comprising the Sepsis data is evident, however the sequences always start with "ERRegistration" and usually terminate with "ReleaseA".

We now discuss the Swiss Health data set results, shown in fig 8, the information value of the KL measure fluctuates between a range of 0.01-0.8. However, in this data entropy has a much higher information content than the other measures and fluctuates between 0.75-1.00. The

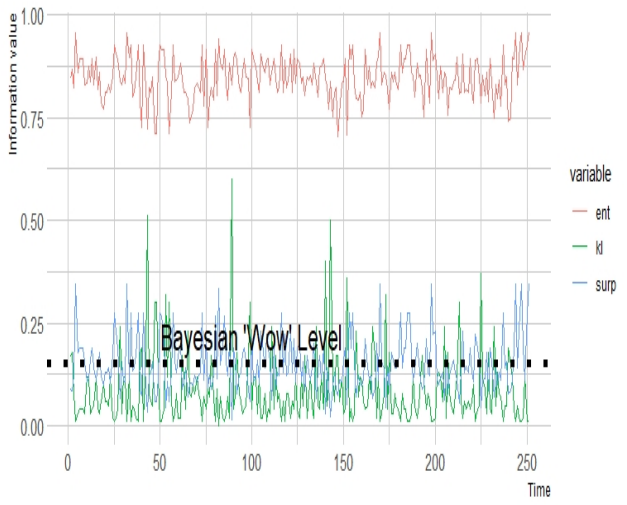

Fig. 7: KL, Entropy and Bayesian Surprise on the Sepsis Medical Data without interest decay

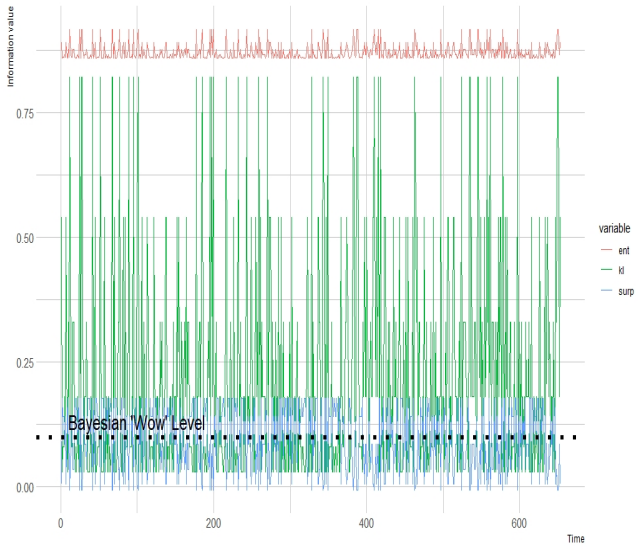

Fig. 8: Entropy, KL and Bayesian Surprise on the Swiss Health data set without interest decay

Bayesian Surprise measure for Sepsis is between 0.1-0.23 and is very different to the characteristics to the KL measure. The "Wow" level is around 0.23. Shown in Fig 5, the equal size and block nature of the sequences comprising the Swiss data is obvious, the sequences are related to Age with younger individuals usually starting with "Very Well" and usually terminate with "Well". The older individuals as we may expect have a more varied response, such as experiencing spates of illness leading to recoveries and wellness again.

Implementing the decay feature, from equation 7, we see for all data sets the rapid decline over time (N) of the Bayesian surprise value. As similar patterns are presented the surprisingness of this data becomes less, finally dropping off to zero as the system has seen it all before. This is similar with typically how humans perceive patterns, as they lose interest in the familiar. In fig 9 the majority of the test patterns become uninteresting after the presentation of about 50 patterns of the total test data of 250 records, however this is the smallest of datasets and corresponds to about 20% of the total test set.

In fig 10 we see the majority of the Swiss Health test patterns become uninteresting after the presentation of about 100 patterns of the usable test data of 650 records, this corresponds to about 15% of the total test set, before similar repeating patterns occur.

The decay function for the biofam test is shown in fig 11 the majority of the test patterns become uninteresting

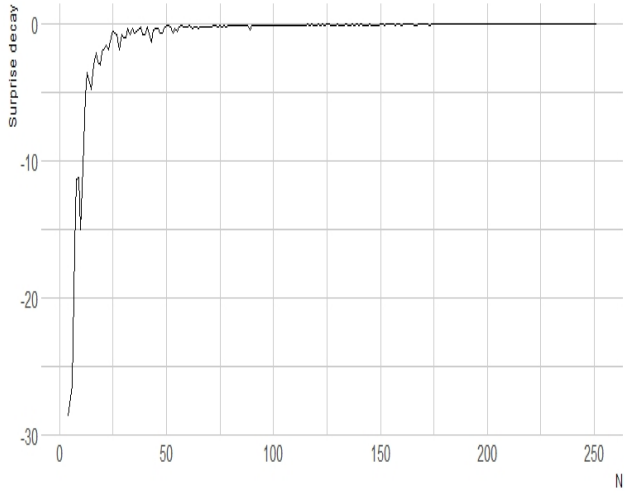

Fig. 9: Bayesian surprise with decay for sepsis data

after the presentation of about 150 patterns of the usable test data of 500 records, this corresponds to about 30% of the total test set. Thereafter, the sequences tend to repeat.

Evaluation of results - statistics

The next stage is to examine the details of the sequences, are there any differences between surprising and non-surprising patterns other than residing in the training

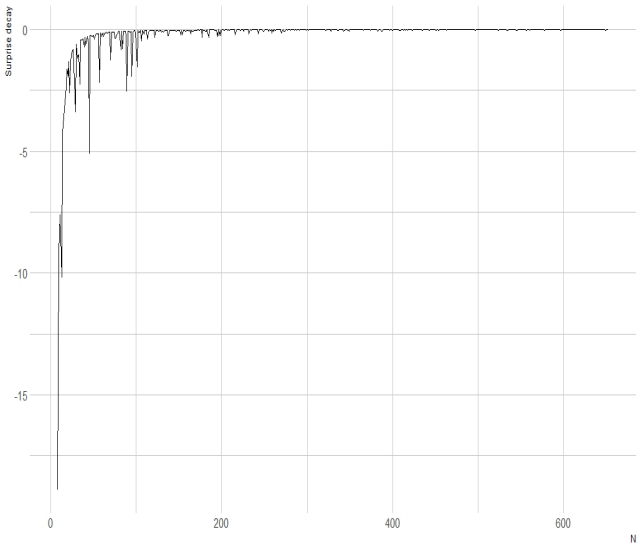

Fig. 10: Bayesian surprise with decay for Swiss Health data

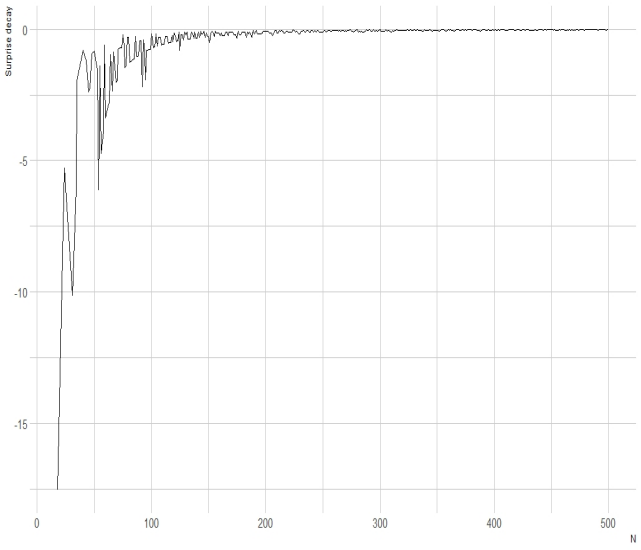

Fig. 11: Bayesian surprise with decay for biofam data

or test set? Referring to table 6 the results for entropy, sequence complexity, longest common prefix and longest common sub-sequence are presented. For the Sepsis sequences the average entropy is 0.74, much larger than Swiss (0.37) and Biofam (0.35) indicating a varied set of symbols. The Turbulence or complexity of the Sepsis data is 11.9 and is more complex than the Swiss (5.3) and Biofam (4.8) data (which is to be expected) given the symbol set. The Longest Common Prefix for the Sepsis data is 20 which is similar to Swiss (20) and smaller than Biofam (32). The Longest Common Sub-sequence for Sepsis data is 10, this is smaller than what could be expected given the long lengths of some sequences. The Swiss data is similar (8) but the Biofam is much larger at 30, this is unusual. Bayesian surprise value is 0.15 for the Sepsis data and is larger than the other two which is to be expected from the wider range of symbols and variable length size of the sequences.

Table 6: Anomaly detection results on the data sets

Dataset	Entropy	Complexity	LCP	LCS	Bayes Surp
Sepsis	0.74	11.9	20	10	0.15
Swiss Health	0.37	5.3	20	8	0.098
BioFam	0.35	4.8	32	30	0.009

Evaluation of results - Optimal Matching Distances (OMD) and Clustering

OMD is an analysis method used to assess the dissimilarity of sequences that usually represent some sort of ordering over time that individuals have experienced (it was developed by social scientists). Clustering is then used on the distance matrices. For the Sepsis OM clustering in fig 12 we required two clusters, however the clustering process did not resolve the differences between surprising and not-surprising sequences, it appears to have clustered based on length of sequences. This is not a useful result and the Biofam and Swiss Health produced much the same null results.

Evaluation of results - Sequence Event Transitions

For all data sets, we tag the sequences with the initial "surprising" and "not-surprising" labels from the test data as it is passed through the Probabilistic Suffix Tree.

In table 7 we highlight the statistically significant events from sub-sequences that discriminate between interesting and not-interesting sub-sequences for the BioFam data. Usually, we are interested in detecting frequently occurring transitions with event sequences. The first column provides the sequence ID, 16 out of 38 sequences are shown. In the second column the full subsequence is shown, creates a distinct (from-state > to-state) event for every discovered transition consisting of a pair of events (end-state event, start-state event) which is assigned to each transition. Similar to association rules these state to state events have a support value based on a minimum support of the number of similar sequences.

Examining the biofam event sub-sequences further in figure 13, we discover the most discriminating sequences between surprising and not-surprising. The chi-square

Table 7: Event sub-sequences discriminate between surprising and not-surprising sequences for biofam data

ID	subseq)	Sup	pvalue	statistic	surprising	not-surprising
1	(Parent)-(Parent>Left)	0.43	0.00	119.53	0.32	0.57
2	(Parent>Left)	0.43	0.00	119.53	0.32	0.57
3	(Parent)-(Parent>Married)	0.12	0.00	37.81	0.16	0.07
4	(Parent>Married)	0.12	0.00	37.81	0.16	0.07
5	(Left>Left+Marr)	0.23	0.00	32.72	0.18	0.29
6	(Parent)-(Left>Left+Marr)	0.23	0.00	32.13	0.18	0.29
7	(Parent)-(Parent>Left)-(Left>Left+Marr)	0.23	0.00	32.13	0.18	0.29
8	(Parent>Left)-(Left>Left+Marr)	0.23	0.00	32.13	0.18	0.29
9	(Parent)-(Parent>Left+Marr)	0.25	0.00	15.80	0.29	0.21
10	(Parent>Left+Marr)	0.25	0.00	15.80	0.29	0.21
11	(Parent)-(Left>Left+Marr)-(Left+Marr>Left+Marr+Child)	0.13	0.02	11.81	0.11	0.16
12	(Parent)-(Parent>Left)-(Left+Marr>Left+Marr+Child)	0.13	0.02	11.81	0.11	0.16
13	(Parent)-(Parent>Left)-(Left>Left+Marr)-(Left+Marr>Left+Marr+Child)	0.13	0.02	11.81	0.11	0.16
14	(Parent>Left)-(Left+Marr>Left+Marr+Child)	0.13	0.02	11.81	0.11	0.16
15	(Parent>Left)-(Left>Left+Marr)-(Left+Marr>Left+Marr+Child)	0.13	0.02	11.81	0.11	0.16
16	(Left>Left+Marr)-(Left+Marr>Left+Marr+Child)	0.13	0.03	11.35	0.11	0.16

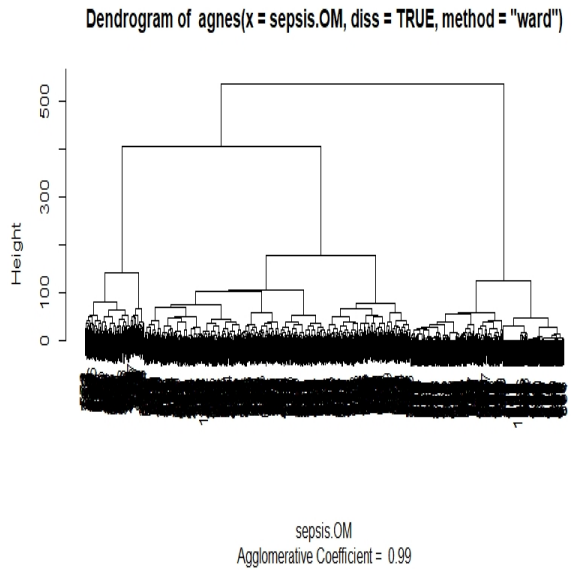

Fig. 12: OM matrix clustered on Sepsis sequences

test is used to test between the two types of sequences, providing a p-value for significance and the Pearson's coefficient, those residuals (darkred) having a value between -1 and 0, the sub-sequence appears *significantly less* frequent than expected under the null hypothesis. Those sequences that are *significantly more* frequent (darkblue) having values between 0 and +1. The surprising patterns have very small values compared with the not-surprising patterns on the 10 most discriminating sequences shown. The not-surprising sequences have values between 0.6 and 0.4.

In table 8 the statistically significant events from the Sepsis data sub-sequences that discriminate between interesting and not-interesting sub-sequences are presented, using the tagged sequences with the surprising label. The Surprising patterns have values for the key variables between 0.6 and 0.9, while the not-surprising sequences events have values much lower between 0.05 and 0.5 - only the Lecoucyte -¿ CRP sequence has a significantly higher value.

The Swiss Health data between surprising and not-surprising test data sub-sequences. In figure 15 the Swiss Health event sub-sequences are shown greater detail, we discover the most discriminating sequences between surprising and not-surprising sequences. The values of the event sequences are generally lower for the surprising patterns with the exception of the "well" event sequence. The not-surprising event sequences generally vary between 0.3 and 0.5 on the Pearson correlation coefficient.

This work only uses the sequences of discrete symbols to detect deviations and novel patterns. It does not use domain knowledge from experts, nor does it use the many other variables provided with each data set. Data sets consisting of variable lengths have historically posed a problem for most probabilistic methods, however the Probabilistic Suffix Tree (PST) and its variable length Markov chain property are ideal for this type of data. In fact variable length sequences such as the Sepsis data are the most interesting in terms of the results. They enable a richer variety of patterns to be encountered. In addition, having many different symbols in the data sets also enables a richer variety of patterns. We find that data sets with small variety of symbols and fixed length sequences such as the Swiss health and Biofam tend not to generate that many interesting patterns and their surprise measure rapidly falls away when the decay parameter is used.

Once trained, the PST outputs a set of probability values for each new input (test) sequence. These new values (one for each symbol in the sequence) are then used as posteriors to be compared with the priors for each symbol based on the trained PST estimates for those symbols. The Bayesian Surprise method will eval-

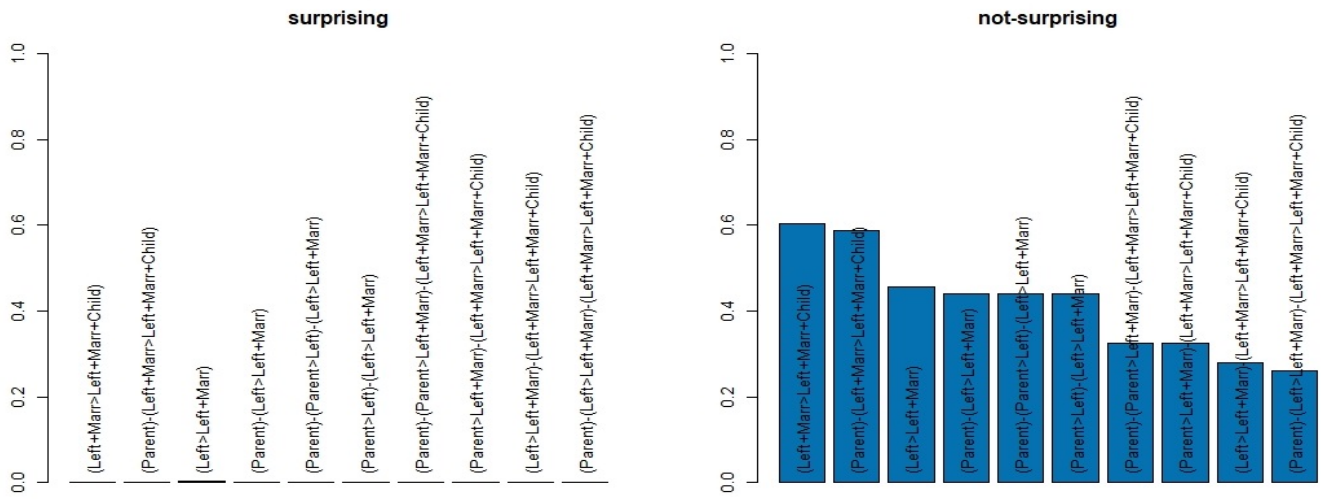

Fig. 13: The 10 most discriminating sub-sequences for the biofam test data

Table 8: Top EVENT sub-sequences discriminate between surprising and not-surprising sequences for Sepsis data

ID	subseq	Sup	pvalue	statistic	surprising	not-surprising
1	(Leucocytes>CRP)-(CRP>Leucocytes)	0.39	0.00	75.48	0.61	0.06
2	(Leucocytes>CRP)-(Leucocytes>CRP)	0.35	0.00	67.87	0.56	0.05
3	(ERRegistration>ERTriage)-(Leucocytes>CRP)-(CRP>Leucocytes)	0.37	0.00	67.86	0.58	0.06
4	(ERRegistration)-(Leucocytes>CRP)-(CRP>Leucocytes)	0.36	0.00	66.39	0.57	0.06
5	(ERRegistration>ERTriage)-(Leucocytes>CRP)-(Leucocytes>CRP)	0.33	0.00	63.75	0.53	0.04
6	(ERRegistration)-(ERRegistration>ERTriage)-(Leucocytes>CRP)-(CRP>Leucocytes)	0.35	0.00	62.11	0.55	0.06
7	(ERRegistration)-(Leucocytes>CRP)-(Leucocytes>CRP)	0.33	0.00	60.82	0.53	0.05
8	(ERRegistration)-(ERRegistration>ERTriage)-(Leucocytes>CRP)-(Leucocytes>CRP)	0.32	0.00	59.67	0.51	0.04
9	(Leucocytes>CRP)	0.73	0.00	58.26	0.91	0.47
10	(CRP>Leucocytes)-(Leucocytes>CRP)	0.37	0.00	58.07	0.57	0.09

Table 9: EVENT sub-sequences discriminate between surprising and not-surprising sequences for Swiss Health data

ID	subseq)	Sup	pvalue	statistic	surprising	not-surprising
1	(very well>well)-(well>very well)	0.40	0.00	50.70	0.27	0.55
2	(very well>well)-(very well>well)	0.34	0.00	40.42	0.22	0.46
3	(very well>well)-(well>very well)-(very well>well)	0.33	0.00	38.79	0.22	0.46
4	(well)	0.54	0.00	37.12	0.65	0.41
5	(very well)-(well>very well)	0.22	0.00	31.96	0.13	0.32
6	(very well)-(very well>well)-(well>very well)	0.22	0.00	31.11	0.13	0.32
7	(so)-(so (average)>well,so)-(so (average))	0.16	0.00	28.54	0.08	0.24
8	(so)-(so (average)>well)-(so (average))	0.16	0.00	28.54	0.08	0.24
9	(very well)-(well>very well)-(very well>well)	0.20	0.00	28.52	0.12	0.29
10	(very well>well)-(very well>well)-(well>very well)	0.19	0.00	28.34	0.11	0.27

uate the differences and determine how anomalous or interesting these patterns are. The decay parameter is essential, otherwise there would be no "memory" of previously observed patterns and the system would continuously view all patterns above the "Wow" cut-off point as interesting. Although, we did not try this, rebuilding the PST's on the test data would be unlikely resolve the issue as we suspect the new patterns would simply change the PST probabilities for the symbols without indicating their true novelty. Therefore, the assessment of novelty has to be a method external to the model used (PST, HMM, etc).

Conclusions

This work contributes to the detection of unusual discrete sequence data that may be of interest to the data miner. Furthermore, we have provided details of why sequence data can be surprising based on several criteria such as entropy, complexity and especially the event sequence transitions. Usually, in outlier detection a patterns interestingness or anomalousness is based on its rarity. Such patterns have low probability and correspond to Shannons original theory. However, using measures that operate on data possessing low probability scores are not useful in practice for identifying outliers

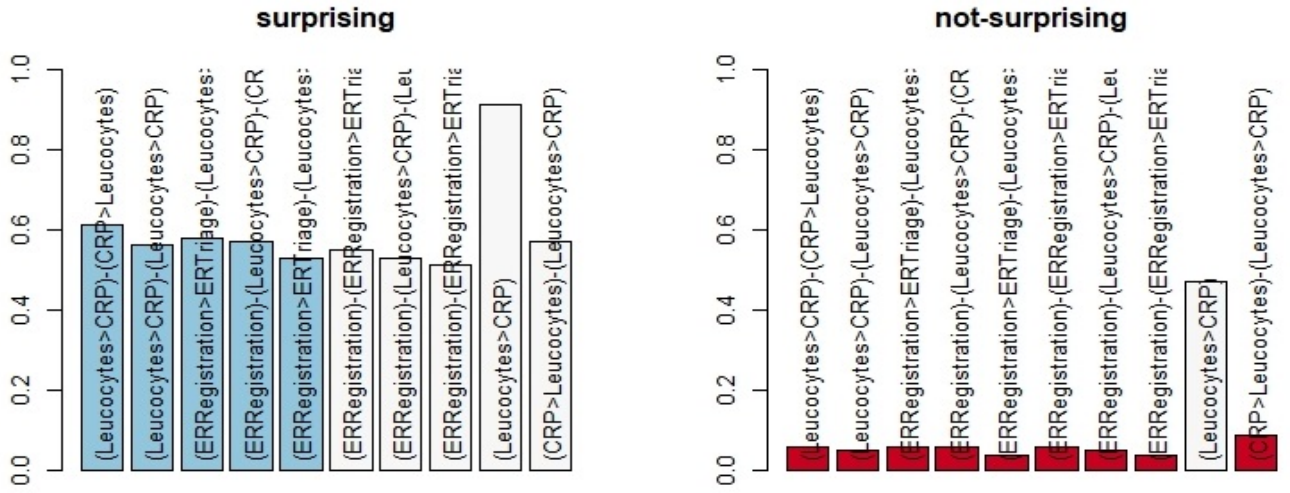

Fig. 14: The 10 most discriminating sub-sequences for Sepsis test data

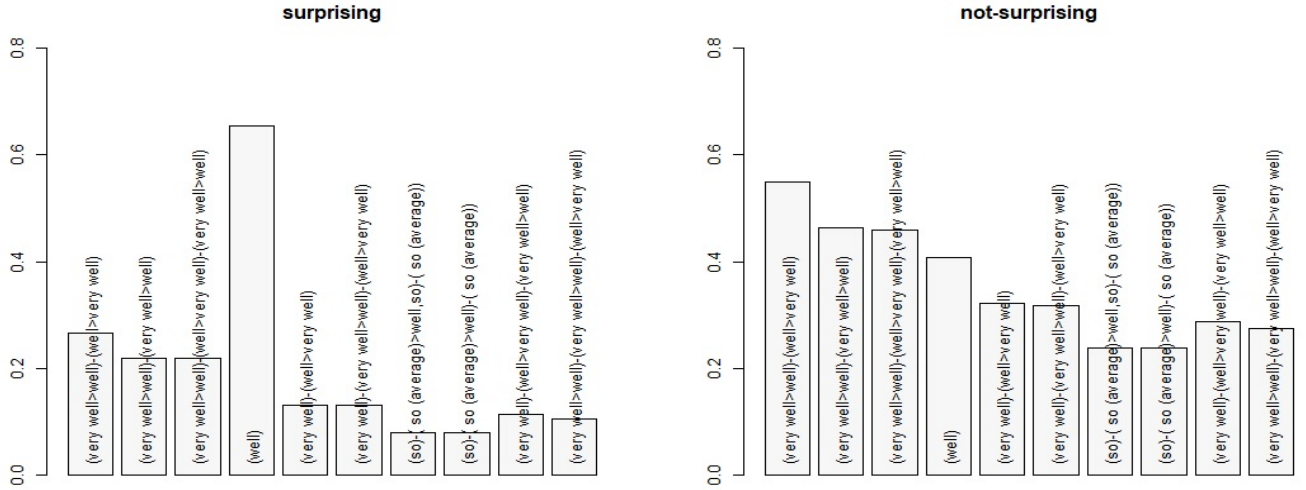

Fig. 15: The 10 most discriminating sub-sequences for Swiss Health test data

(interesting patterns), as all such patterns would be deemed interesting. Bayesian surprise can detect data with other properties that would be missed by information theoretic measures such as Shannon surprise and entropy for example. Bayesian surprise is the relative entropy (difference) between the prior and posterior distributions, when identical surprise is zero, when differences are large surprise is large.

Undoubtedly, better information theoretic measures or combinations of measures could be employed to detect variations/similarities between discrete symbol sequences. However, in this work we achieved our aims of developing a plausible, human-like method of reacting to interesting or unusual patterns and also implementing the loss of interest as we encounter more of the same patterns. However, the surprising patterns discovered by our approach may or may not useful to the data miner, the system is self-consistent based on the similarity on previous patterns. Future work will include domain knowledge (where available) from experts to guide and constrain the Bayesian surprise method and the decay function. It would be of value to use the associated variables such as demographics for the Swiss data and results of medical tests for the Sepsis data. We may consider the use of Deep Neural learning such as Recurrent Networks to analyze sequences replacing the Probabilistic Suffix Tree.

Compliance with Ethical Standards

Conflict of interests

We declare that we have no conflicts of interest.

Ethical Approval

This article does not contain any studies with human participants or animals performed by any of the authors.

Acknowledgments

We would like to thank Satu Helske for advice on the variable length, hidden markov model approach.

References

1. Abbott A (1995) Sequence analysis: New methods for old ideas. Annual Review of Sociology 21(1):93–113, DOI 10.1146/annurev.soc.21.1.93
2. Avila L, Martínez E (2014) Behavior monitoring under uncertainty using bayesian surprise and optimal action selection. Expert Systems with Applications 41(14):6327 – 6345
3. Bain R (2016) Are our brains bayesian? Significance 13(4):14–19, DOI https://doi.org/10.1111/
4. Baldi P, Itti L (2010) Of bits and wows: A bayesian theory of surprise with applications to attention. Neural Networks 23:649–666, DOI 10.1016/j.neunet.2009.12.007
5. Barto A, Mirolli M, Baldassarre G (2013) Novelty or surprise? Frontiers in Psychology 4:907, DOI 10.3389/fpsyg.2013.00907
6. Bayarri MJ, Morales J (2003) Bayesian measures of surprise for outlier detection. Journal of Statistical Planning and Inference 111:3–22
7. Becattini N, Borgianni Y, Cascini G, Rotini F (2017) Surprise and design creativity: investigating the drivers of unexpectedness. International Journal of Design Creativity and Innovation 5(1-2):29–47, DOI 10.1080/21650349.2015.1090913
8. Berchtold A (2010) Sequence analysis and transition models. In: Breed M, Moore J (eds) Encyclopedia of Animal Behavior, Academic Press, Oxford., pp 139–145, DOI 10.1016/b978-0-08-045337-8.00233-3
9. Berger J, Bernardo J, Sun D (2009) The formal definition of reference priors. The Annals of Statistics 2(37):905–938
10. Berlyne D (1994) Conflict, Arousal, and Curiosity. Oxford University Press
11. Boldt M, Borg A, Ickin S, Gustafsson J (2019) Anomaly detection of event sequences using multiple temporal resolutions and markov chains. Knowledge and Information Systems 62:669–686, DOI 10.1007/s10115-019-01365-y
12. Bulmann P, Wyner A (1999) Variable length markov chains. The Annals of Statistics 27(3):480–513, DOI 10.1214/aos/1018031204
13. Cohen P, Heeringa B, Adams N (2002) Unsupervised segmentation of categorical time series into episodes. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings., pp 99–106, DOI 10.1109/ICDM.2002.1183891
14. Correll M, Heer J (2017) Surprise! bayesian weighting for de-biasing thematic maps. IEEE Transactions on Visualization and Computer Graphics 23(1):651–660
15. Ekman P, Davidson R (1960) The nature of emotion: Fundamental questions. McGraw-Hill
16. Elzinga C (2010) Complexity of categorical time series. Sociological Methods & Research 38(3):463–481, DOI 10.1177/0049124109357535
17. Elzinga C, Studer M (2015) Spell sequences, state proximities, and distance metrics. Sociological Methods & Research 44(1):3–47, DOI 10.1177/0049124114540707
18. Florez-Larrahondo G, Bridges S, Vaughn R (2005) Efficient modeling of discrete events for anomaly detection using hidden markov models. Information Security 3650:506–514
19. Gabadinho A, Ritschard G (2016) Analyzing state sequences with probabilistic suffix trees: The PST R package. Journal of Statistical Software 72(3):1–39, DOI 10.18637/jss.v072.i03
20. Gabadinho A, Ritschard G, Müller N, Studer M (2011) Analyzing and visualizing state sequences in R with TraMineR. Journal of Statistical Software 40(4):1–37, DOI 10.18637/jss.v040.i04
21. Gottlieb J, Oudeyer P, Lopes M, Baranes A (2013) Information-seeking, curiosity, and attention: computational and neural mechanisms. Trends Cognitive Science 17:585–593, DOI doi:10.1016/j.tics.2013.09.001.
22. Grassi P, Bartels A (2021) Magic, Bayes and wows: A Bayesian account of magic tricks. Neuroscience and Biobehavioral Reviews 126:515–527, DOI https://doi.org/10.1016/j.neubiorev.2021.04.001
23. Griffiths TL, Tenenbaum J (2006) Optimal predictions in everyday cognition. Psychological Science 17(9):767–773, DOI https://doi.org/10.1111/j.

1740-9713.2016.00935.x
24. Hasanbelliu E, Kampa K, Príncipe J, Cobb JT (2012) Online learning using a bayesian surprise metric. The 2012 International Joint Conference on Neural Networks (IJCNN) pp 1–8
25. Huang J, Jaysawal B, Wang C (2021) Mining full, inner and tail periodic patterns with perfect, imperfect and asynchronous periodicity simultaneously. Data Mining and Knowledge Discovery 35:1225–1257
26. Itti L, Baldi P (2005) A principled approach to detecting surprising events in video. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol 1, pp 631–637 vol. 1, DOI 10.1109/CVPR.2005.40
27. Itti L, Baldi P (2009) Bayesian surprise attracts human attention. Vision Research 49(10):1295 – 1306, DOI https://doi.org/10.1016/j.visres.2008.09.007
28. Keogh E, Lonardi S, Chiu B (2002) Finding surprising patterns in a time series database in linear time and space. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, p 550–556
29. Kullback S, Leibler R (1951) On information and sufficiency. The Annals of Mathematical Statistics 1(22):79–86
30. Largeron-Leténo C (2003) Prediction suffix trees for supervised classification of sequences. Pattern Recognition Letters 24(16):3153–3164, DOI https://doi.org/10.1016/j.patrec.2003.08.002
31. Lee M, Wagenmakers E (2013) Bayesian Cognitive Modeling. Cambridge University Press
32. Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Briefings in Bioinformatics 11(5):473–483
33. Liao T, Fasang A (2021) Comparing groups of life-course sequences using the bayesian information criterion and the likelihood-ratio test. Sociological Methodology 51:44–85
34. Lin J, Keogh E, Wei L, Lonardi S (2007) Experiencing sax: A novel symbolic representation of time series. Data Mining Knowledge Discovery 15(2):107–144
35. Maguire P, Moser P, Maguire R, Keane M (2019) Seeing patterns in randomness: A computational model of surprise. Topics in Cognitive Science 11(1):103–118
36. Marsland S (2003) Novelty detection in learning systems. Neural Comp Surveys 3:1–39
37. McGarry K (2005) A survey of interestingness measures for knowledge discovery. Knowledge Engineering Review 20(1):39–61
38. Needleman S, Wunsch C (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453
39. Nevill-Manning C, Witten I (1997) Identifying hierarchical structure in sequences: A linear-time algorithm. J Artif Int Res 7(1):67–82
40. de Oliveira H, Ospina R (2018) A note on the shannon entropy of short sequences. CoRR abs/1807.02603, URL http://arxiv.org/abs/1807.02603, 1807.02603
41. Rabiner L, Juang B (1986) An introduction to hidden markov models. IEEE ASSP Magazine 3(1):4–16, DOI 10.1109/MASSP.1986.1165342
42. Rhienberger C, Hammitt J (2018) Dinner with bayes: On the revision of risk beliefs. Journal of Risk and Uncertainty 57(3):253–280
43. Rieck K, Laskov P (2008) Linear-time computation of similarity measures for sequential data. Journal Machine Learning Research 9:23–48
44. Ritschard G (2021) Measuring the nature of individual sequences. Sociological Methods & Research 0(0), DOI 10.1177/00491241211036156
45. Schmidhuber J (2010) Formal theory of creativity, fun, and intrinsic motivation (1990-2010). IEEE Trans Auton Ment Dev 2(3):230–247, URL http://dblp.uni-trier.de/db/journals/tamd/tamd2.html#Schmidhuber10
46. Statisticat, LLC (2020) LaplacesDemon: Complete Environment for Bayesian Inference. URL https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software, r package version 16.1.4
47. Wilson W, Birkin P, Aickelin U (2007) The motif tracking algorithm. International Journal of Automation and Computing 5(1):32–44, DOI 10.1007/s10453-004-5872-7
48. Yang P, Chen K, Ching-Chi H (2018) Subjective Association Rule Mining: From Point-based Ranking Sequence to Interval-based Temporal Sequence. In: Proceedings of the 10TH International Conference on Machine Learning and Computing (ICMLC 2018), Assoc Computing Machinery, 1515 Broadway, New York, NY 10036-9998 USA, pp 167–171
49. Zalewski W, Silva F, Wu F, Lee H, Maletzke A (2012) A symbolic representation method to preserve the characteristic slope of time series. Springer-Verlag, Berlin, Heidelberg, SBIA'12, p 132–141
50. Zolfaghari S, Khodabandehloo E, Riboni D (2021) TraMiner: Vision-Based Analysis of Locomotion Traces for Cognitive Assessment in Smart-Home. Cognitive Computation pp 626–688, DOI 10.1007/s12559-020-09816-3