Anomaly Detection and Explanation in Time-Series Data

Hajar Homayouni, Sudipto Ghosh, Indrakshi Ray

Department of Computer Science, Colorado State University

Hajar. Homayouni@colostate.edu

Data Quality Tests

Validate data in a data store to detect violations of constraints that are imposed by application domain experts and data model

- **Constraints over single attributes**

  patient_weight>=0

- **Constraints over multiple attributes**

  pregnancy_status=true → patient_gender=female

- **Constraints over multiple records**

  patient_weight growth rate over time must be positive and in the range [4 lb, 22 lb] for every infant

Time-Series Data

Time series T is a sequence of d-dimensional records

$T = \langle R_1, \ldots, R_n \rangle$

d=1 for univariate time series
d>1 for multivariate time series

where

- $R_i = (R_i^1, \ldots, R_i^d)$ is the value of time series at time i
- Each dimension corresponds to an attribute

Different Anomalies in Time Series

- **Anomalous records:** Given a time series T, a faulty record R_t is one that its observed value is significantly different from expected value of T at t
- **Anomalous sequences:** Given a set of time series $T = \{T_1, \ldots, T_m\}$, an anomalous sequence $T_j \in T$ is one whose behavior is significantly different from majority of time series in T

Limitations of Existing Approaches

Stochastic modeling techniques

- Are appropriate for univariate time series data
- Do not consider non-linear associations among records
- Assume that data follows a known statistical distribution

Machine learning modeling techniques

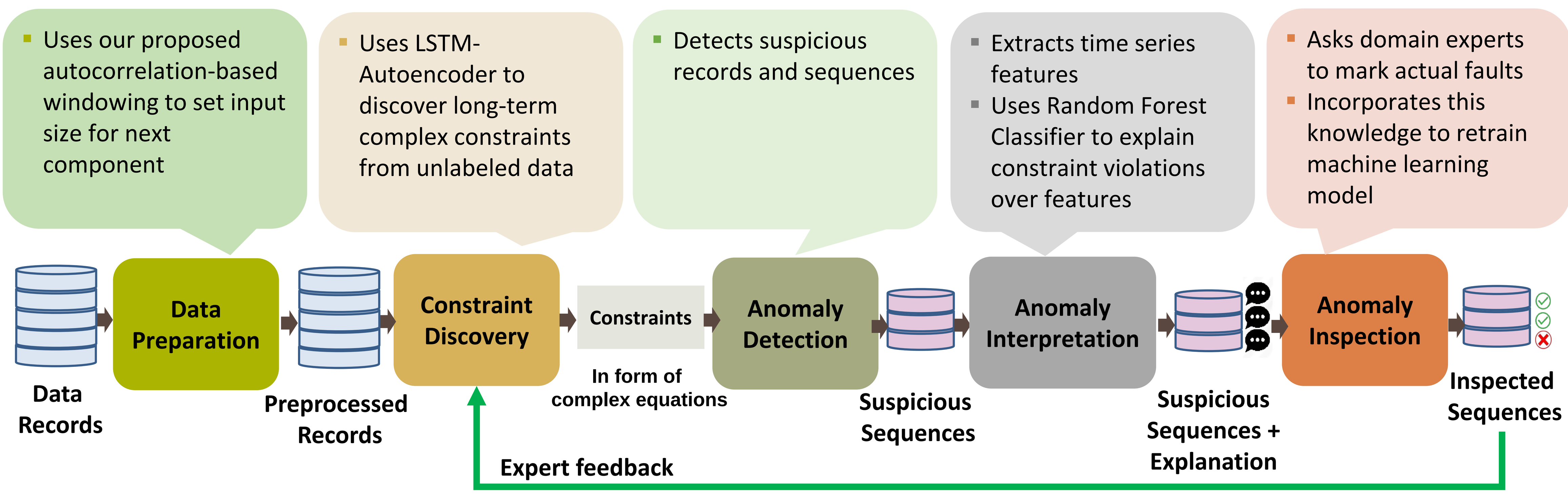
- Have potential to generate false alarms
- Do not explain constraint violations

Research Goal

Develop an automated data quality test approach that

- Discovers constraints in complex long-term associations among records and attributes
- Reports as suspicious sequences and records that do not satisfy constraints
- Determines constraints that are violated by suspicious records and sequences
- Minimizes false alarms over time through an interactive process

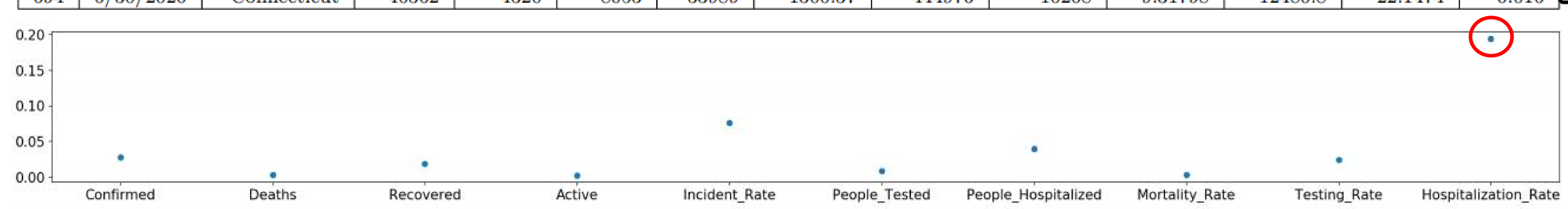
IDEAL: Interactive Detection and Explanation of Anomalies using an LSTM-Autoencoder

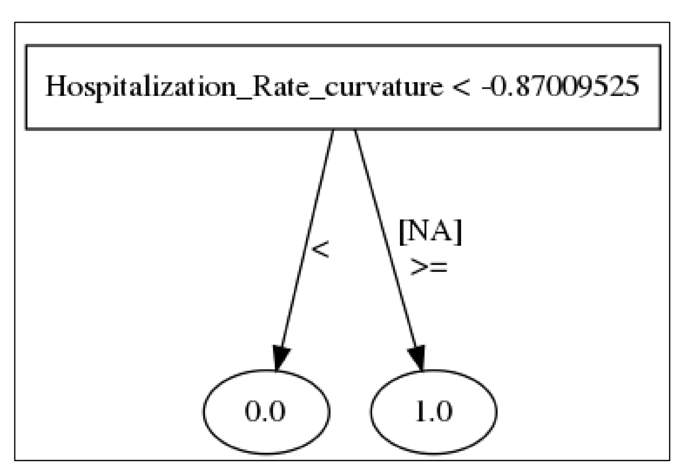


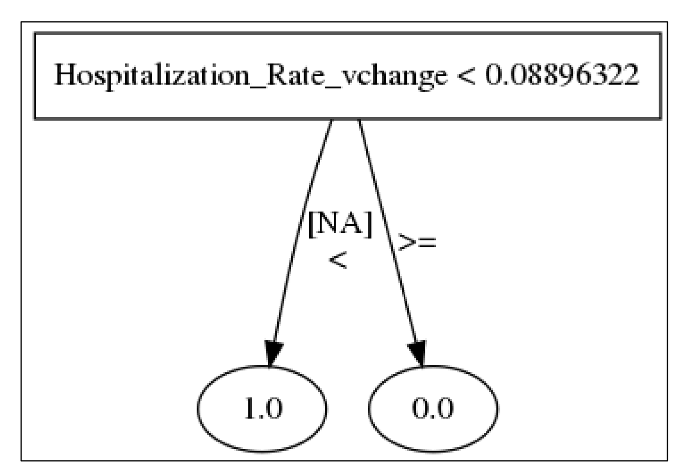
Suspicious Sequence Detected from Johns Hopkins COVID-19 Data

Sequence 1

id	time	Province State	Con-firmed	Deaths	Re-cov-ered	Active	Incident Rate	People Tested	People Hospi-talized	Mortal-ity Rate	Testing Rate	Hospi-taliza-tion Rate	s-score
616	4/12/2020	Connecticut	12035	554	nan	11481	337.56	41220	1654	4.60324	1156.15	13.7432	0.006
...	...	...	...	...	...	...	...	...	...	...	...	...	...
656	5/23/2020	Connecticut	39640	3637	7127	28876	1111.83	202328	12538	9.17508	5674.94	31.6297	0.019
...	...	...	...	...	...	...	...	...	...	...	...	...	...
668	6/4/2020	Connecticut	43091	3989	7511	31591	1208.63	276436	13793	9.25715	7753.54	32.009	0.021
...	...	...	...	...	...	...	...	...	...	...	...	...	...
694	6/30/2020	Connecticut	46362	4320	8053	33989	1300.37	444976	10268	9.31798	12480.8	22.1474	0.016

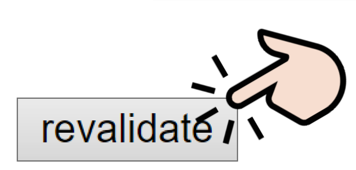






Constraints

- If curvature(*Hospitalization_Rate*)= NA, OR curvature(*Hospitalization_Rate*)≥ -0.87, then suspicious subsequence is invalid.
- If vchange(*Hospitalization_Rate*)= NA, OR vchange(*Hospitalization_Rate*)< 0.08, then suspicious subsequence is invalid.



Mutation Analysis

Objective: Inject *mutants*, which are faulty records or sequences that mimic typical anomalies in time-series data

Mutation Operator	Description
*M*1–Add noise	Adds random noise to an attribute of randomly selected records from entire dataset.
*M*2–Horizontal shift	Shifts attribute values of records in a subset of records along time axis.
*M*3–Vertical shift	Adds a random value to all attribute values in a subset of records.
*M*4–Re-scale	Multiplies all the attribute values in a subset of records with a random number.
*M*5–Add dense noise	Changes all attribute values in a subset of records to randomly selected values.

Evaluation Goals

We used mutated datasets to evaluate three aspects of IDEAL:

1. Constraint discovery and anomaly detection effectiveness
2. Anomaly explanation effectiveness
3. Performance

P_t: number of actual faulty sequences
TP_t: number of actual faulty sequences detected as suspicious by the tool
N_t: number of actual valid sequences
FP_t: number of valid sequences incorrectly detected as suspicious by the tool

Metrics

TT: total time it takes to perform automated steps of IDEAL

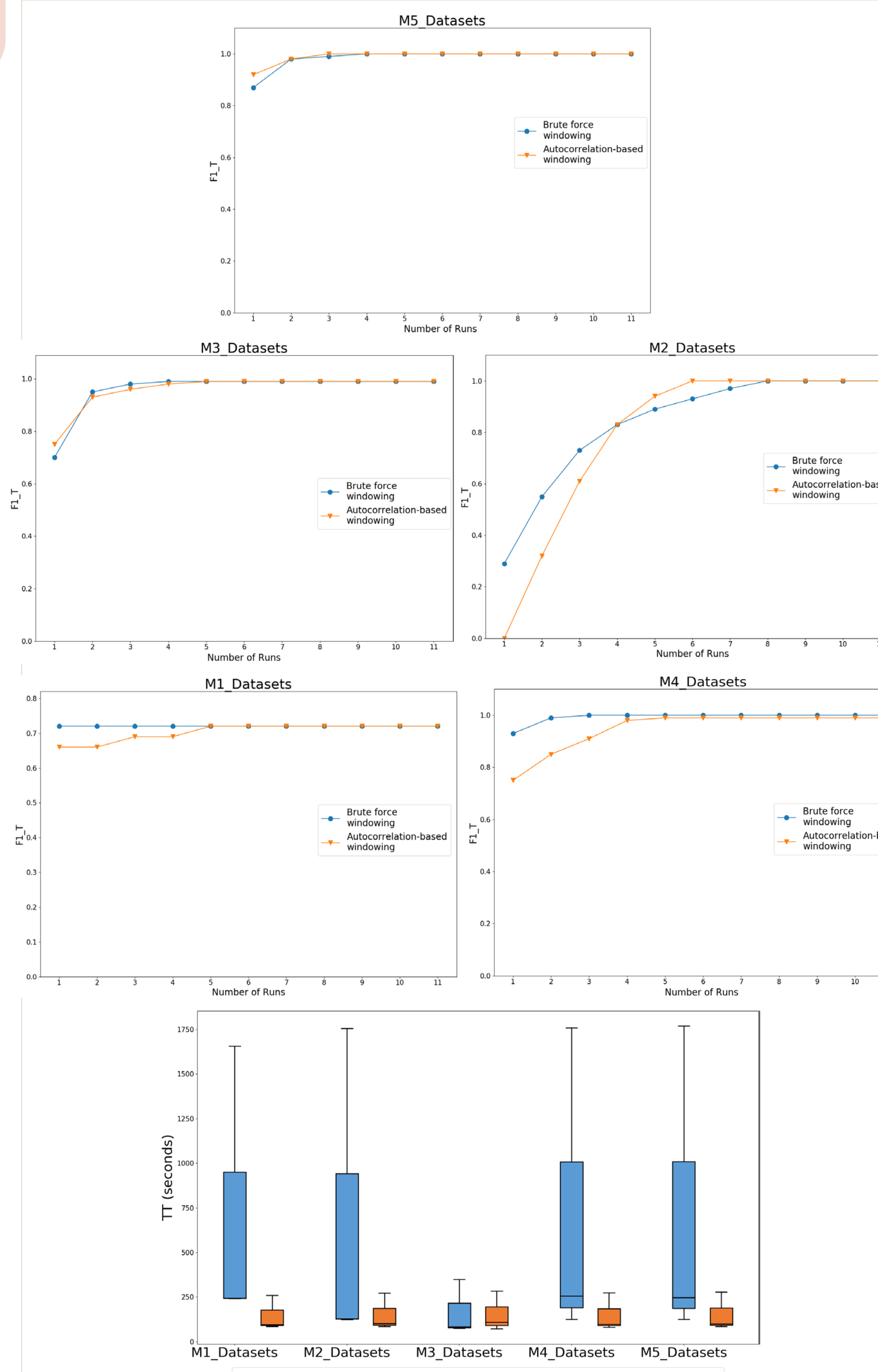
$$F1_t = 2 \times \frac{(precision_t \times recall_t)}{(precision_t + recall_t)}$$

$$precision_t = \frac{TP_t}{(TP_t + FP_t)}, recall_t = TPR_t, FPR_t = \frac{FP_t}{N_t}, TPR_t = \frac{TP_t}{P_t}$$

Subjects

- NASA Shuttle datasets (58,000 records, 7 attributes)
- Yahoo servers traffic datasets (1,420 records, 1 attribute)

Results



Conclusions

- Developed a deep-learning based approach to discover long-term complex constraints in unlabeled data
- Added explanation to the detected suspicious sequences
- Detected different types of anomalies which we created using mutation analysis
- Minimized false alarms in an interactive process

IDEAL

1) Detected 35% to 75% of injected faults in its first execution
2) Accuracy of approach improved over time
3) Autocorrelation-based windowing was almost as effective but 3.88 times more efficient than brute force windowing
4) Visualization plots could correctly explain constraints violated by suspicious sequences

Future Work

- Anomaly detection in non-primitive and mixed data types
- Anomaly detection in streaming data
- Adversarial sampling for training and validating anomaly detection techniques

This research is supported by grants from

- US National Science Foundation
- Google Cloud, AFRL, Cable Labs, and Furuno Electric Company