Factorizing Event Sequences

Naren Sundaravaradan and Naren Ramakrishnan, *Virginia Tech*
David A. Hanauer, *University of Michigan Medical School*

Factorizing interleaved event sequences, such as those found in electronic medical records, into simpler processes can yield new insights from large datasets.

Large datasets often include sequences of events. For example, electronic medical records (EMRs) can be viewed as sequences of ICD-9 and CPT-4 codes. ICD-9 (International Classification of Diseases, ninth revision) is a coding system for injuries, diseases, and other health-related conditions, while CPT-4 (Current Procedural Terminology, fourth edition) is a categorization system for medical procedures such as surgeries and lab tests.

In EMRs, it's common for such coding sequences to be interleaved—thus, a patient's heart history is intermingled with x-ray reports for a sports-related injury, or a plastic surgery procedure is mixed in with a record of kidney disease.

While one subsystem of the body can have complicating side effects on another subsystem, there's a need to factorize event sequences across a patient population into nonredundant processes to discover clinically relevant patterns.

EVENT SEQUENCE FACTORIZATION

Event sequence factorization draws on both process and sequence mining. Process mining uses temporal data to reconstruct a process—represented, say, by a Kripke structure or Petri net—that could account for it. Sequence mining aims to identify patterns in sequences that recur frequently or that optimize a user-defined objective.

There is an inherent tradeoff between mining local patterns, which are more efficient to mine and can be quite detailed, and global patterns, which yield more succinct representations. Our approach, which focuses on separating interleaved event sequences, finds the "sweet spot" between these two options. While mining local data, event sequence factorization doesn't generate an excessive number of patterns, but at the same time it can define a global model of the underlying processes.

Figure 1 shows two simple examples of event sequence factorization.

In Figure 1a, the factor sequences RABCD and EWXYZD generate the sequence AWXBYCDZ. Every element in the generated sequence must come from one of the given factors. Note that an element, such as D, can occur in multiple processes. Also, there must be an order-preserving mapping from a subset of factor elements to the generated sequence. However, factorization restricts these subsets

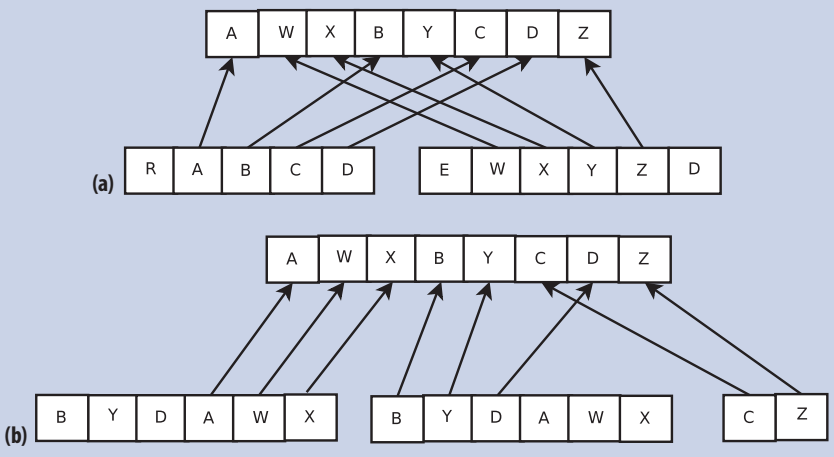

Figure 1: Event sequence factorization. (a) The factor sequences RABCD and EWXYZD generate the sequence AWXBYCDZ. (b) Factorization with two applications of the same factor.

0018-9162/12/$31.00 © 2012 IEEE Published by the IEEE Computer Society

Kidneys

(D584.9: renal failure, acute nitric oxide synthase [NOS]) -> (DV42.0: transplant, kidney) -> (D585.6: renal disease, end stage) -> (D414.00: coronary atherosclerosis of unspecified vessel) -> (D577.1: pancreatitis, chronic)

Leading to a psychotherapy session

(D256.9: dysfunction, ovarian NOS) -> (D628.9: infertility, female NOS) -> (P90806: psychotherapy, office, 45-50 min.) -> (D309.9: reaction, adjustment NOS)

(D611.9: disorder, breast NOS) -> (P76091: mammogram, both breasts) -> (D611.72: lump or mass in breast) -> (P76090: mammogram, one breast) -> (D311: disorder, depressive, not elsewhere classified [NEC]) -> (P90844: psychotherapy, individual, 45-50 min.)

(D310.1: personality change due to clubbing, cyanosis, and edema [CCE]) -> (P90812: interactive psychotherapy, office, 45-50 min.) -> (D294.9: disorder, persistent mental, due to CCE/NOS) -> (D300.3: disorder, obsessive-compulsive) -> (P90806: psychotherapy, office, 45-50 min.)

Brain

(D345.01: epilepsy, generalized nonconvulsive, with intractable epilepsy) -> (P95819: EEG, awake and asleep) -> (D345.10: epilepsy, generalized convulsive) -> (P70553: MRI brain with and without dye) -> (D191.9: neoplasm, malignant, brain NOS)

(P70553: MRI brain with and without dye) -> (DV67.2: chemotherapy follow-up) -> (D191.3: neoplasm, malignant, brain, parietal lobe) -> (P96413: chemotherapy, IV infusion, 1 hr.)

Psoriasis

(D696.0: psoriatic arthritis) -> (D696.1: psoriasis NEC) -> (P99222: initial hospital care) -> (D244.9: hypothyroidism NOS) -> (D250.00: diabetes mellitus)

(D696.1: psoriasis NEC) -> (D250.02: type 2 diabetes) -> (D457.1: lymphedema NEC) -> (D250.00: diabetes mellitus) -> (P99222: initial hospital care)

Pelvis

(P73510: x-ray exam of hip) -> (D715.95: osteoarthrosis NOS, pelvis/thigh) -> (DV43.64: hip joint replacement status) -> (D733.90) -> (P81000: urinalysis, nonautomated, with scope) -> (DV42.0: transplant, kidney) -> (P72170: x-ray exam of pelvis)

(D617.3: endometriosis, pelvic peritoneum) -> (D628.9: infertility, female NOS) -> (P76857: ultrasound exam, pelvic, limited) -> (D256.9: dysfunction, ovarian NOS)

Liver

(D996.82: complication, liver transplant) -> (P99141: sedation, conscious, IV, intramuscular, Isoniazid) -> (P75984: x-ray control catheter change) -> (P74305: x-ray bile ducts/pancreas) -> (P47525: change bile duct catheter) -> (P47505: injection for liver x-rays)

Lungs

(DV42.6: transplant, lung) -> (D996.84: complication, transplanted lung) -> (D792.9: abnormal finding, body substance NEC) -> (D512.8: pneumothorax, spontaneous NEC) -> (D212.3: neoplasm, benign, bronchus/lung) -> (P88312: special stains)

(D496: obstruction, chronic airway NEC) -> (P94720: monoxide diffusing capacity) -> (P94360: measure airflow resistance) -> (P94240: residual lung capacity) -> (P94060: evaluation of wheezing)

(P99283: emergency dept. visit) -> (D493.90: asthma) -> (D530.1: esophagitis NOS) -> (D518.82: insufficiency, pulmonary NEC) -> (D518.3: eosinophilia, pulmonary) -> (D493.91: asthma NOS with status asthmaticus) -> (P94010: breathing capacity test) -> (P94720: monoxide diffusing capacity) -> (P94240: residual lung capacity)

Others

(D427.41: fibrillation, ventricular) -> (P93737: analyze cardioverter-defibrillator without reprogramming) -> (D185: neoplasm, malignant, prostate) -> (D715.91: osteoarthrosis NOS, shoulder) -> (P20610: drain/inject, joint/bursa)

(D250.01: diabetes mellitus, uncomplicated, type 1) -> (D240.9: goiter NOS) -> (P76536: ultrasound exam of head and neck) -> (D784.2: swelling in head/neck)

(D696.1: psoriasis NEC) -> (D571.2: cirrhosis, alcoholic, liver) -> (D696.0: psoriatic arthropathy)

Figure 2. Processes discovered by factorizing event sequences in an electronic medical record dataset.

to form a contiguous subsequence—in this case, ABCD in the first factor sequence and WXYZ in the second. Note that some factor elements, such as R, need not appear in the generated sequence.

In Figure 1b, one of the factor sequences, BYDAWX, is repeated. Again, note the order-preserving mapping and the contiguity of subsequences AWX and BYD.

Given a particular sequence, it's easy to derive many factorizations from it. But given a large database of sequences, the goal is to derive a small set of processes that can generate all of the sequences in the database.

SEQUENCE FACTORIZATION ALGORITHM

Details of the sequence factorization algorithm we developed are beyond the scope of this article, but it's essentially incremental. We construct and maintain a model by adding one process at a time. As the algorithm encounters a dataset

sequence, it maintains a working set of processes, each in one of four states: *waiting*—the process hasn't yet been factorized; *converging*—the process has been factorized but hasn't yielded sufficient evidence to make a decision; *converged*—the process is chosen to be included in the model (over its factorization); and *invalid*—the process is no longer needed but can't be removed because another process depends on it. As the descriptions indicate, a process traverses the states in order: waiting, converging, converged, and possibly invalid.

Because the algorithm incrementally computes a factorization, the resulting model isn't optimal. Nevertheless, it significantly compresses EMR data. Furthermore, it has revealed several patterns about medical diagnoses and procedures.

EXPERIMENTAL RESULTS

With approval from the University of Michigan Health System, we organized a dataset of de-identified information from about 1.6 million patients who received care there. The actual medical records contained about 100 million time-stamped ICD-9 and CPT-4 codes.

We ran our algorithm on three sets of 150,000 patients with an alphabet size of 10,000 for each set. To further condense the representation of patient records, we collapsed a sequence of contiguous identical events into one event—thus, AABBCCBD became ABCBD.

As with most large-scale studies involving the discovery of clinical associations, we manually reviewed a subset of the data to determine significant and interesting patterns. Figure 2 shows a sampling of results from our analysis.

Many of the processes we discovered were consistent with known medical information.

For example, ventricular fibrillation (Dx 427.41) -> automatic implantable cardioverter-defibrillator check (Px 93737) -> malignant prostate cancer (Dx 185) -> shoulder osteoarthrosis (Dx 715.91) -> shoulder joint injection (Px 20610) is a process that might be found in an elderly man. Similarly, the medical history of a patient with an autoimmune disorder might include type 1 diabetes (Dx 250.01) -> goiter (Dx 240.9) -> neck ultrasound (Px 76536) -> neck swelling (Dx 784.2). Goiter is often associated with autoimmunethyroiditis, and type 1 diabetes is also an autoimmune disorder.

Other processes revealed unfortunate stories about patients.

For example, intractable seizures (Dx 345.01) -> EEG, brain scan (Px 96819) -> brain MRI, to look for pathology (Px 70553) -> malignant brain neoplasm (Dx191.9) indicates that a patient had severe seizures and, after a workup to determine the cause, was found to have a brain tumor. Another example is ovarian dysfunction (Dx 256.9) -> female infertility (Dx 628.9) -> psychotherapy (Px 90806) -> adjustment reaction (Dx309.9), which is a psychiatric diagnosis defining a significant emotional response to a specific stressor—in this case, being unable to bear children.

Some patterns were clinically interesting but are less well known in the medical domain. For example, we discovered an association between psoriasis and hypothyroidism that has been documented but is rare: psoriatic arthritis (Dx 696.0) -> psoriasis (Dx 696.1) -> hospitalization (Px 99222) -> hypothyroidism (Dx 244.9) -> diabetes mellitus (Dx 250.00).

Another interesting pattern with psoriasis involves lymphedema, which is swelling caused by blockage of the lymphatic system: psoriasis (Dx 696.1) -> type 2 diabetes (Dx 250.02) -> lymphedema (Dx 457.1) -> diabetes mellitus (Dx 250.00) -> hospitalization (Px 99222). This condition has also been reported in the medical literature.

Both patterns include diabetes, which might be expected given that psoriasis and diabetes are both associated with elevated body mass index and obesity.

Finally, some of the temporal patterns we discovered are quite complex, such as acute renal failure (Dx 584.9) -> kidney transplant (Dx V42.0) -> end stage renal disease (Dx 585.6) -> coronary atherosclerosis (Dx 414.00) -> chronic pancreatitis (Dx 577.1). Acute (as opposed to chronic) pancreatitis is a known cause of acute renal failure, but this process suggests that chronic renal failure—akin to end stage renal disease—might cause pancreatitis.

We've developed a novel approach to factorizing events sequences into a small set of processes, and have demonstrated its effectiveness in deriving insights from EMR data. A major advantage of our approach is that it can be used in a distributed data mining setting, making it ideal for mining remote databases as well as for when privacy preservation is important. Another benefit is that this approach combines local and global considerations of pattern mining.

Naren Sundaravaradan *is a PhD student in the Department of Computer Science at Virginia Tech. Contact him at narens@vt.edu.*

Naren Ramakrishnan*, Discovery Analytics column editor, is the Thomas L. Phillips Professor of Engineering in the Department of Computer Science and director of the Discovery Analytics Center at Virginia Tech. Contact him at naren@vt.edu.*

David A. Hanauer, *MD, is assistant director, Comprehensive Cancer Center Bioinformatics Core, University of Michigan Medical School. Contact him at hanauer@umich.edu.*

Editor: Naren Ramakrishnan, Dept. of Computer Science, Virginia Tech, Blacksburg, VA; naren@cs.vt.edu