Novelty Detection in Patient Histories: Experiments with Measures Based on Text Compression

Ole Edsberg, Øystein Nytrø and Thomas Brox Røst

Department of Computer and Information Science (IDI)

also associated with: Norwegian Centre for EHR Research (NSEP)

Norwegian University of Science and Technology (NTNU)

NTNU
Norwegian University of
Science and Technology

The problem

- Patient histories can contain hundreds of encounter notes.
- The information in the encounter notes is often redundant. Examples: status summaries, follow-ups of established treatment.
- Clinicians and other people need to read / scan patient histories to uncover the important information, but have very limited time.

Our goal

Develop a method to pick out the encounter notes describing *novel* developments in a patient history. If successful, this would enable us to provide novelty-based highlighting, folding, summarization etc in the patient record interface.

Our test case

General practice, where the problem is especially prevalent.

Data model

A patient history is a sequence of encounters. Each encounter has:

- A textual note
- Some diagnosis codes, such as “K86 Hypertension uncomplicated” or “L93 Tennis Elbow”
- Possibly other information (prescriptions, lab tests . . .).

Our approach

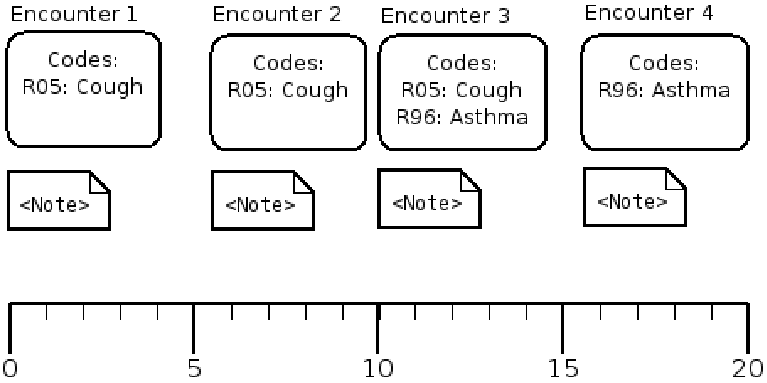
Define measures that rate the novelty of a note by comparing it to all previous notes in the history.

(Data other than the note could be taken into account, but we leave that for later work.)

Evaluation Criterion

- Ideally, we would have clinicians manually rate the novelty of a large number of notes. However, this would be expensive and would involve subjectivity in exactly how *novelty* should be defined.
- In this work, we instead labeled a note as novel iff it had any associated diagnosis codes that did not occur earlier in the history. This gave us 273991 classification examples.
- For each candidate novelty measure, we calculate a novelty rating for each note, draw the ROC curve, and use the area under the curves (AUC) as a performance score.

Example



Encounter 1

Codes:
R05: Cough

<Note>

Encounter 2

Codes:
R05: Cough

<Note>

Encounter 3

Codes:
R05: Cough
R96: Asthma

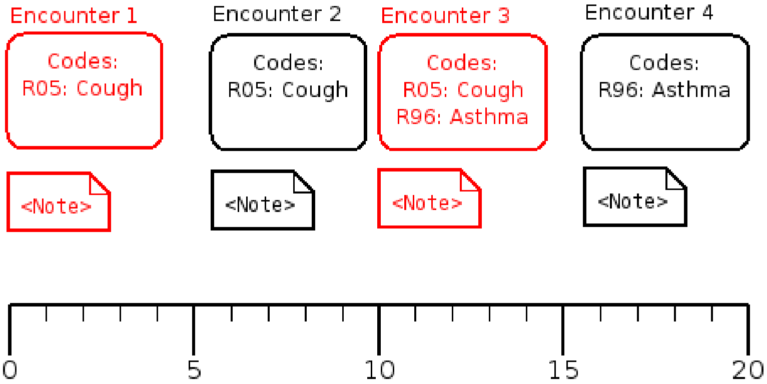
<Note>

Encounter 4

Codes:
R96: Asthma

<Note>

Example


Encounter 1
Codes:
R05: Cough
<Note>
Encounter 2
Codes:
R05: Cough
<Note>
Encounter 3
Codes:
R05: Cough
R96: Asthma
<Note>
Encounter 4
Codes:
R96: Asthma
<Note>
0
5
10
15
20

Methods

We tested the following candidate novelty measures:

- Four measures based on text compression (as an approximation of Kolmogorov complexity).
- A TF-IDF vector space measure transplanted from related work on news bulletins.
- Six trivial baselines.
- SVM combining the above.

Preliminaries: Kolmogorov Complexity

- $K(x)$, the Kolmogorov complexity of the bit string x, is *the length of the shortest binary program with no input that makes a fixed universal Turing machine output* x. It quantifies the algorithmic information content in x. In essence, it is the size of the ultimate compression of x.
- $K(x)$ cannot be computed, but we can upper-bound it with $C(x)$, the size of the compression of x with some other compression algorithm.
- $K(x|y)$, the conditional Kolmogorov complexity of x given y, is *the length of the shortest binary program that makes the machine output* x *when given* y *as input*.
- $K(xy)$ is the Kolmogorov complexity of the concatenation of x and y.

Preliminaries: Information Distance

- [Bennet *et al.* 1998] proposed the Information Distance, $\mathrm{ID}(x,y)$, which is *the length of the shortest program that makes the machine output x when given y and vice versa.* They expressed it in terms of Kolmogorov complexity:

$$\mathrm{ID}(x,y) = \max\{K(x|y), K(y|x)\}.$$

- A normalized version is:

$$\mathrm{NID}(x,y) = \frac{\max\{K(x|y), K(y|x)\}}{\max\{K(x), K(y)\}}.$$

- ID and NID are metrics and have been proved to be minimal among a wide set of admissible distances (with a density requirement). (Both properties up to an additive constant.)
- They are not computable, of course.

NCD, the normalized compression distance

- [Li *et al.* 2004], [Keogh *et al.* 2004] and [Cilibrasi and Vitányi 2005] use compression algorithms (`gzip` etc) to approximate Kolmogorov complexity and apply real-world analogues of $\mathrm{NID}(x,y)$ to different similarity-based tasks, with good results.
- [Cilibrasi and Vitányi 2005]'s formulation, the normalized compression distance:

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}.$$

- We used the NCD (with `gzip` as the compressor C) in two of our candidate measures.

Our proposal: Normalized Asymmetric Compression Distance (NACD)

- The NCD (repeated from previous slide) is symmetric:

$$\mathrm{NCD}(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}.$$

- Novelty detection (a comparison between the previous notes and the current note) is inherently asymmetric. Therefore, we propose an asymmetric variant of the NCD:

$$\mathrm{NACD}(x,y) = \frac{C(xy) - C(x)}{C(y)}$$

  (This is an approximation of $\frac{K(y|x)}{K(y)}$.)

- We used the NACD (with `gzip` as the compressor C) in another two of our candidate measures.

How to use these distances to rate novelty

Given a distance D and a sequence of notes $h = (n_1, n_2, \ldots, n_n)$ and a goal of rating the novelty of note k, we tried two alternatives:

Comparing the concatenation of the previous notes to the current note

$$\text{CONCAT-NOVELTY}(h,k) = \mathrm{D}(\underset{i\in[1,k)}{\text{concat}}(n_i), n_k)$$

Comparing each previous note to the current note and taking the minimum

$$\text{MIN-NOVELTY}(h,k) = \min_{i\in[1,k)} \mathrm{D}(n_i, n_k)$$

This gives rise to four novelty measures: NCD-CONCAT, NCD-MIN, NACD-CONCAT and NACD-MIN.

TF-IDF vector space measure

- Related work: Detection of the first coverage of each new topic in a stream of news bulletins.
- We transplated their baseline measure, which is the cosine distance between incremental TF-IDF (Term Frequency $\times$ Inverse Document Frequency) weighted vector representations of the documents.

Trivial measures

Some reality checks for our other measures:

- The rank of the note in the history.
- The day of the note relative to the beginning of the history
- Then number of days since the previous note
- The number of characters in the note
- The number of previously unseen words in the note
- The compressed size of the note.

Combined measure

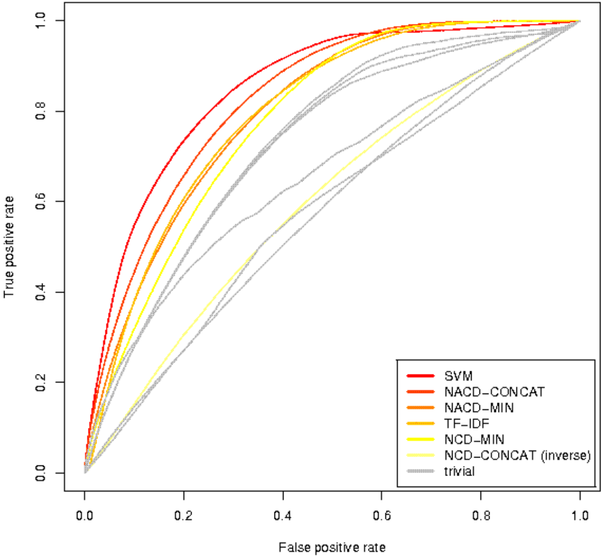
We used all the other measures as input to an SVM (radial basis kernel, default parameters). (Leaving out any one measure did not improve performance.)

Results: Areas under the ROC curves

Novelty measure	Area under ROC curve
SVM of all others	0.850
NACD-CONCAT	0.826
NACD-MIN	0.798
TF-IDF vector space	0.795
NCD-MIN	0.780
Trivial: compressed size of note	0.735
Trivial: number of characters	0.727
Trivial: number of new words	0.722
Trivial: days since last	0.656
NCD-CONCAT (inverse)	0.600
Trivial: days since beginning	0.576
Trivial: rank of note (inverse)	0.575

(We took the inverse of some measures that gave $AUC < 0.5$.)

Results: The ROC curves



Conclusions

- Our measures were to some extent able to detect novel notes.
- Performance was too low for these measures to stand alone as a clinical tool. However, it seems likely that they can be useful as part of a greater tool considering more types of information.
- Surprise: NCD between current and concatenated previous notes performed badly.
- Discovery: NACD between current and concatenated previous notes was the best performer except for SVM. (We haven't seen anybody else consider this type of measure for novelty detection.)

Future Work

Planned:

- Have physicians mark up novelty themselves. (Will not give enough data to learn from, but is needed for better evaluation.)
- Use text-based measures as a part of a bigger system also using other parts of the patient record.

Also possible:

- Test on news bulletin data.
- Use more sophisticated NLP in the TF-IDF measure.
- Sentence level novelty detection.

Thanks for listening!