Insider Threat Detection Through Anomalies in Electronic Medical Records

You Chen

Department of Biomedical Informatics
Vanderbilt University Medical Center
Five key components of data processing

• Data Representation
• Data Normalization
• Data Discretization
• Similarity Measurements
• Dimensionality Reduction
Data Representation

The real world

Information in the real world are abstract and cannot be directly modeled

Data

Data which can be used by computational models

Data are collected by mapping entities in the domain of interest to symbolic representation by means of some measurement procedure, which associates the value of a variable with a given property of an entity.

An example of data representation in access logs of EHR

Various data representations

• Flatted data
  – Data matrix or table
• Time series
• Text
• Image and video
• Workflow data
Distinct Patient Accesses across Time

Accesses in week days

Accesses in weekend

Time

Number of Patients

0 5000 10000 15000 20000 25000 30000

Subjective

Chief Complaint
Headaches

Patient History
- Illnesses - asdfsdfsdfsdfsdfsds; Measles asdf
- Operations - Appendectomy - 1987
- Social History - Married; Employed
- Family History - Denies All

Allergies
- Uncategorized - Peanut-Allergen-Ingredient
- Uncategorized - Advil Liqui-Gel-Allergen-Medication
- Uncategorized - Augmentin-Allergen-Medication
- Uncategorized - Amoxicillin-Allergen-Ingredient
- Uncategorized - Tylenol-Allergen-Medication
- Uncategorized - Vytorin 10-10-Allergen-Medication - Difficulty breathing/constricting of throat, Dizziness, Congestion, Anxiety, Confusion, Drowsiness, Intolerance; comments
- Uncategorized - Dust allergy (disorder)
- Uncategorized - Allergy to animal hair (disorder)

Current Medications
- Vicodin; Date: 01/01/2008; Sig:
**Images in EHR**

### Images in EHR

#### Chest X-ray 2 Views

**Preliminary**

Images available for viewing. Please click on the link above.

**Order Details**

- **Order by:** Provider: Test
- **Reviewed by:** Radiology North
- **Order:** #17596
- **Overview:** chestX-ray
- **Ordered by:** Provider: Test
- **Supervisor:** Provider: Test
- **Approval:** Not Required

---

**Results History**

**Chest X-ray 2 Views**

- **Report 1:** 15/08/2022 14:41:01 PM - Chest X-ray 2 Views
- **Report 2:** 15/08/2022 14:41:01 PM - Chest X-ray 2 Views

**Addendum:**

- This is an addendum to the tasking
- Addendum: Details on tasking
- Result to check on tasking

Images available for viewing. Please click on the link above.
Clinical Workflow
Types of attribute scales

• Nominal Scale
• Ordinal Scale
• Numerical Scale
  – Ratio Scale
  – Interval Scale
Nominal Scale

• The values of the attribute are only “labels”, which is used to distinguish each other
  – Finite number of values
  – No order information
  – No algebraic operation could be conducted, except those related to frequency
  – An example
    – \{1,2,3\} -> \{doctor, nurse, pharmacist\}
      -> \{Medical Information Service, Clinical Trials Center, Breast Center\}
A table describing accesses of roles and users on a patient across time

<table>
<thead>
<tr>
<th>Accessed Role</th>
<th>Number of Accesses</th>
<th>Number of Users</th>
<th>Time Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>doctor</td>
<td>10</td>
<td>3</td>
<td>t1-t2</td>
</tr>
<tr>
<td>nurse</td>
<td>20</td>
<td>5</td>
<td>t2-t3</td>
</tr>
<tr>
<td>pharmacist</td>
<td>15</td>
<td>6</td>
<td>t3-t4</td>
</tr>
<tr>
<td>doctor</td>
<td>16</td>
<td>4</td>
<td>t4-t5</td>
</tr>
<tr>
<td>nurse</td>
<td>22</td>
<td>5</td>
<td>t5-t6</td>
</tr>
<tr>
<td>nurse</td>
<td>6</td>
<td>2</td>
<td>t6-t7</td>
</tr>
</tbody>
</table>

Nominal Attribute

# of providers

![Bar chart showing the number of providers over roles: doctors, nurses, and pharmacists.](chart.png)
Ordinal Scale

• The values of the attribute is to indicate certain ordering relationship resided in the attribute
  – Order is more important than value
  – No algebraic operation could be conducted, except those related to sorting

• An example
  – Richter’s scale on earthquake
A heartquake of 5.5 magnitudes is more important than one of 3 but less than one 9. Which indicates that there is an order between the data.

<table>
<thead>
<tr>
<th>Richter magnitudes</th>
<th>Description</th>
<th>Earthquake effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 2.0</td>
<td>Micro</td>
<td>Micro earthquakes, not felt.</td>
</tr>
<tr>
<td>2.0-2.9</td>
<td>Minor</td>
<td>Generally not felt, but recorded.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Often felt, but rarely causes damage.</td>
</tr>
<tr>
<td>3.0-3.9</td>
<td>Light</td>
<td>Noticeable shaking of indoor items, rattling noises. Significant damage unlikely.</td>
</tr>
<tr>
<td>4.0-4.9</td>
<td>Moderate</td>
<td>Can cause major damage to poorly constructed buildings over small regions. At most slight damage to well-designed buildings.</td>
</tr>
<tr>
<td>5.0-5.9</td>
<td>Strong</td>
<td>Can be destructive in areas up to about 160 kilometers (100 mi) across in populated areas.</td>
</tr>
<tr>
<td>6.0-6.9</td>
<td>Major</td>
<td>Can cause serious damage over larger areas.</td>
</tr>
<tr>
<td>7.0-7.9</td>
<td>Great</td>
<td>Can cause serious damage in areas several hundred miles across.</td>
</tr>
<tr>
<td>8.0-8.9</td>
<td></td>
<td>Devastating in areas several thousand miles across.</td>
</tr>
<tr>
<td>9.0-9.9</td>
<td>Epic</td>
<td>Never recorded; see below for equivalent seismic energy yield.</td>
</tr>
<tr>
<td>10.0+</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Despite having the same interval of 0.5, the difference from one point to another in the scale in Joule is not uniform.

<table>
<thead>
<tr>
<th>Richter Approximate Magnitude</th>
<th>Joule equivalent</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>63.1 kJ</td>
</tr>
<tr>
<td>0.5</td>
<td>355 kJ</td>
</tr>
<tr>
<td>1.0</td>
<td>2.00 MJ</td>
</tr>
<tr>
<td>1.5</td>
<td>11.2 MJ</td>
</tr>
<tr>
<td>2.0</td>
<td>63.1 MJ</td>
</tr>
<tr>
<td>2.5</td>
<td>355 MJ</td>
</tr>
<tr>
<td>3.0</td>
<td>2.00 GJ</td>
</tr>
<tr>
<td>3.5</td>
<td>11.2 GJ</td>
</tr>
</tbody>
</table>
Numerical Scale

• The values of the attribute is to indicate quantity of some predefined unit
  – There should be a basic unit, which can be transferred to another one
  – The value is how many copies of the basic unit
  – Some algebraic operations could be conducted

• Two types of numerical scale
  – Interval scale
  – Ratio scale
Differences of Ratio Scale and Interval Scale

• An interval variable is a measurement where the difference between two values is meaningful.
  – The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees.

• A ratio variable, has all the properties of an interval variable, and also has a clear definition of 0.0. When the variable equals 0.0, there is none of that variable.
  – Variables like number of accesses on a patient, number of accesses of a user are ratio variables.
  – Temperature, expressed in F or C, is not a ratio variable. A temperature of 0.0 on either of those scales does not mean 'no temperature'.
  – However, temperature in Kelvin is a ratio variable, as 0.0 Kelvin really does mean 'no temperature'.
<table>
<thead>
<tr>
<th>Accessed Role</th>
<th>Number of Accesses</th>
<th>Number of Users</th>
<th>Temperature (F)</th>
<th>Temp (C)</th>
<th>Time Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>docotr</td>
<td>10</td>
<td>3</td>
<td>100.2</td>
<td>38</td>
<td>t1-t2</td>
</tr>
<tr>
<td>nurse</td>
<td>20</td>
<td>5</td>
<td>99.3</td>
<td>37</td>
<td>t2-t3</td>
</tr>
<tr>
<td>pharmacist</td>
<td>15</td>
<td>6</td>
<td>99.1</td>
<td>37</td>
<td>t3-t4</td>
</tr>
<tr>
<td>doctor</td>
<td>16</td>
<td>4</td>
<td>98.2</td>
<td>37</td>
<td>t4-t5</td>
</tr>
<tr>
<td>nurse</td>
<td>22</td>
<td>5</td>
<td>97.5</td>
<td>36</td>
<td>t5-t6</td>
</tr>
<tr>
<td>nurse</td>
<td>6</td>
<td>2</td>
<td>97.8</td>
<td>36</td>
<td>t6-t7</td>
</tr>
</tbody>
</table>

**Ratio Variable**

14.8333 Mean  
6.01383 STD

**Interval Variable**

98.68333 Mean  
1.026483 STD

**Covariance**

5.86111
### Which Computing Operations Could be Done?

<table>
<thead>
<tr>
<th>Algebraic Operation</th>
<th>Nominal</th>
<th>Ordinal</th>
<th>Interval</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency distribution</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Median and Percentiles</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Add or Subtract</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mean, Standard Deviation</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Ratio, or Coefficient of Variation</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
What we need to know to conduct access logs auditing?

- Data Representation
- Data Normalization
- Data Discretization
- Similarity Measurements
- Dimensionality Reduction
Why we need data normalization?

Min-Max Normalization

(a) Euclidean distance before normalization

(b) Euclidean distance after normalization
Three types of normalization

• The attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0, 0.0 to 1.0
  – Min-max normalization
  – Z-score normalization
Min-Max
Normalization

- Performs a linear transformation on the original data
- Support: \( \text{min}_A \) and \( \text{max}_A \) are the minimum and maximum values of an attribute, \( A \).
- Min-max normalization maps a value, \( v \), of \( A \) to \( v' \) in the range \([\text{new\_min}_A, \text{new\_max}_A]\) by computing:

\[
v' = \frac{v - \text{min}_A}{\text{max}_A - \text{min}_A} (\text{new\_max}_A - \text{new\_min}_A) + \text{new\_min}_A
\]
An Example of Min-Max Normalization

• Let *income range $12,000 to $98,000 normalized to [0.0, 1.0]*.

• Then $73,600 is mapped to

\[
\frac{73,600 - 12,000}{98,000 - 12,000} \times (1.0 - 0) + 0 = 0.716
\]

\[
\text{new}_{\text{max}_A} \quad \text{new}_{\text{min}_A}
\]

\[
\text{max}_A \quad \text{min}_A
\]
Z-Score Normalization

• Change the original data quite a bit
• The values for an attribute, A, are normalized based on the mean (A) and standard deviation (A) of A.
• A value, v, of A is normalized to v’ by computing:

\[ v' = \frac{v - \bar{A}}{\sigma_A} \]
distance is represented by units of standard deviations from the mean

<table>
<thead>
<tr>
<th>Accessed Role</th>
<th>Number of Accesses</th>
<th>Number of Users</th>
<th>Temperature (F)</th>
<th>Temp (C)</th>
<th>Time Span</th>
</tr>
</thead>
<tbody>
<tr>
<td>doctor</td>
<td>10</td>
<td>3</td>
<td>100.2</td>
<td>38</td>
<td>t1-t2</td>
</tr>
<tr>
<td>nurse</td>
<td>20</td>
<td>5</td>
<td>99.3</td>
<td>37</td>
<td>t2-t3</td>
</tr>
<tr>
<td>pharmacist</td>
<td>15</td>
<td>6</td>
<td>99.1</td>
<td>37</td>
<td>t3-t4</td>
</tr>
<tr>
<td>doctor</td>
<td>16</td>
<td>4</td>
<td>98.2</td>
<td>37</td>
<td>t4-t5</td>
</tr>
<tr>
<td>nurse</td>
<td>22</td>
<td>5</td>
<td>97.5</td>
<td>36</td>
<td>t5-t6</td>
</tr>
<tr>
<td>nurse</td>
<td>6</td>
<td>2</td>
<td>97.8</td>
<td>36</td>
<td>t6-t7</td>
</tr>
</tbody>
</table>

The role nurse during time t2-t3 has accesses above average—a distance of 0.8591 above the average accesses.
What we need to know to conduct access logs auditing?

- Data Representation
- Data Normalization
- Data Discretization
- Similarity Measurements
- Dimensionality Reduction
Data Discretization

• Dividing the range of a continuous attribute into intervals
• Interval labels can then be used to replace actual data values
• Reduce the number of values for a given continuous attribute
Why Data Discretization?

• Many learning methods—like association rules, Bayesian networks can handle only discrete attributes

• This leads to a concise, easy-to-use, knowledge-level representation of mining results

The goal of discretization is to reduce the number of values for a continuous attribute

grouping them into a number, n, of intervals (bins).
Discretization

Illustration of the supervised vs. unsupervised discretization
Binning

• The sorted values are distributed into a number of buckets, or bins, and then replacing each bin value by the bin mean or median

• Binning is a top-down splitting technique based on a specified number of bins

• Binning is an unsupervised discretization technique, because it does not use class information
Two Methods of Binning

• Equal-width (distance) partitioning
  – Divides the range into N intervals of equal size
  – if A and B are the lowest and highest values of the attribute, the width of intervals will be: \( W = (B - A) / N \)
  – The most straightforward, but outliers may dominate presentation

• Equal-depth (frequency) partitioning
An Example of Equal-width Partitioning

- Sorted data for price (in dollars):
  4, 8, 15, 21, 21, 24, 25, 28, 34
- \[ W = \frac{(B - A)}{N} = \frac{(34 - 4)}{3} = 10 \]
  Bin 1: 4-14, Bin 2: 15-24, Bin 3: 25-34
- Equal-width (distance) partitioning:
  Bin 1: 4, 8
  Bin 2: 15, 21, 21, 24
  Bin 3: 25, 28, 34
Distribution of users in hospital on two measurements

Equal Width Partitioning

Size of the network is 6

Average dependent relation for the network

Size of 1-nearest neighbor network

Distribution of users in hospital on two measurements

Size of the network is 6
• Equal-depth (frequency) partitioning
  – Divides the range into N intervals, each containing approximately the same number of samples
  – Good data scaling

• Example
  – Sorted data for price (in dollars):
    • 4, 8, 15, 21, 21, 24, 25, 28, 34
  – Equal-depth (frequency) partitioning:
    • Bin 1: 4, 8, 15
    • Bin 2: 21, 21, 24
    • Bin 3: 25, 28, 34
• Data Representation
• Data Normalization
• Data Discretization
• Similarity Measurements
• Dimensionality Reduction
Why care about similarity?

• Represent the internal relationship between data objects

• It is essential to many data mining algorithms
Distance Measurements

• Distance measure can be used to characterize the concept of “similarity”

• Distance or Metric should satisfy
  – Non-negativity: \[ d(i, j) \geq 0 \text{ and } d(i, j) = 0 \text{ iff } i = j \]
  – Symmetry \[ d(i, j) = d(j, i) \text{ for all } i, j \]
  – Triangle inequality \[ d(i, j) \leq d(i, k) + d(k, j) \text{ for all } i, j \text{ and } k \]
Minkowski Distance

- Minkowski Distance is a generalization of Euclidean Distance

\[ dist = \left( \sum_{k=1}^{d} |p_k - q_k|^{r} \right)^{\frac{1}{r}} \]

Where \( r \) is a parameter, \( d \) is the number of dimensions (attributes) and \( p_k \) and \( q_k \) are, respectively, the kth attributes (components) of data objects \( p \) and \( q \).
• $r = 1$. City block (Manhattan, taxicab, $L_1$ norm) distance

• $r = 2$. Euclidean distance

• $r \rightarrow \infty$. “supremum” ($L_{\max}$ norm, $L_{\infty}$ norm) distance
  – This is the maximum difference between any component of the vectors
Euclidean Distance

- Euclidean Distance

\[
dist = \sqrt{\sum_{k=1}^{d} (p_k - q_k)^2}
\]

Normalization is necessary, if scales differ.
Euclidean Distance-Examples

Dis(p1,p2)=\sqrt{(0-2)^2+(2-0)^2}=\sqrt{8}=2.828

<table>
<thead>
<tr>
<th>point</th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>p1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>p2</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>p3</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>p4</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

Distance Matrix

\[
\begin{array}{cccc}
  p1 & p2 & p3 & p4 \\
  p1 & 0 & 2.828 & 3.162 & 5.099 \\
  p2 & 2.828 & 0 & 1.414 & 3.162 \\
  p3 & 3.162 & 1.414 & 0 & 2 \\
  p4 & 5.099 & 3.162 & 2 & 0 \\
\end{array}
\]
Weighted Minkowski distance

\[ d(x_1, x_2) = \left( \sum_{k=1}^{d} w_k (x_1(k) - x_2(k))^p \right)^{\frac{1}{p}} \]

*Reflects the importance of each attribute*

In both weighted and unweighted Minkowski distance, each attribute contribute independently to the measure of distance
Mahalanobis Distance

• Mahalanobis distance standardizes data not only in the direction of each attributes but also based on the covariance between attributes

\[
\text{mahalanobis} (p, q) = \sqrt{(p - q) \sum^{-1} (p - q)^T}
\]

Where \(p\) and \(q\) are two data points in \(d\) dimensions

\(\Sigma\) is the covariance matrix of the input data \(X\), the size of it is \(d\) by \(d\). “\(d\)” is the number of attributes or variables

\[
\Sigma_{j,k} = \frac{1}{n-1} \sum_{i=1}^{n} (X_{ij} - \overline{X}_j)(X_{ik} - \overline{X}_k)
\]
An Example of Mahalanobis Distance for Two Data Points

Step 1: There are three data points in two attributes
A: (0.5, 0.5);
B: (0, 1);
C: (1.5, 1.5)

Step 2: Covariance Matrix Calculation for two variables

\[
\Sigma = \begin{bmatrix}
0.58 & 0.25 \\
0.25 & 0.25 \\
\end{bmatrix}
\]

\[
\Sigma^{-1} = \begin{bmatrix}
3 & -3 \\
-3 & 7 \\
\end{bmatrix}
\]

Step 3: Distance calculation

\[
\text{Mahal}(A, B) = \begin{bmatrix}
0.5 - 0 \\
0.5 - 1 \\
\end{bmatrix} \Sigma^{-1} \begin{bmatrix}
0.5 - 0 \\
0.5 - 1 \\
\end{bmatrix} = 4
\]

\[
\text{mahalanobis}(p, q) = (p - q) \Sigma^{-1} (p - q)^T
\]
Similarity Between Binary Vectors

• Common situation is that objects, $p$ and $q$, have only binary attributes

• Compute similarities using the following quantities

  \[ M_{01} = \text{the number of attributes where } p \text{ was 0 and } q \text{ was 1} \]
  \[ M_{10} = \text{the number of attributes where } p \text{ was 1 and } q \text{ was 0} \]
  \[ M_{00} = \text{the number of attributes where } p \text{ was 0 and } q \text{ was 0} \]
  \[ M_{11} = \text{the number of attributes where } p \text{ was 1 and } q \text{ was 1} \]

• Simple Matching and Jaccard Distance/Coefficients

  \[ \text{SMC} = \frac{\text{number of matches}}{\text{number of attributes}} = \frac{(M_{11} + M_{00})}{(M_{01} + M_{10} + M_{11} + M_{00})} \]

  \[ J = \frac{\text{number of value-1-to-value-1 matches}}{\text{number of not-both-zero attributes values}} = \frac{(M_{11})}{(M_{01} + M_{10} + M_{11})} \]
SMC versus Jaccard: Example

\[ p = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \]
\[ q = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \]

\[ M_{01} = 2 \quad \text{(the number of attributes where } p \text{ was 0 and } q \text{ was 1)} \]
\[ M_{10} = 1 \quad \text{(the number of attributes where } p \text{ was 1 and } q \text{ was 0)} \]
\[ M_{00} = 7 \quad \text{(the number of attributes where } p \text{ was 0 and } q \text{ was 0)} \]
\[ M_{11} = 0 \quad \text{(the number of attributes where } p \text{ was 1 and } q \text{ was 1)} \]

SMC = \((M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00}) = (0 + 7) / (2 + 1 + 0 + 7) = 0.7\)

J = \((M_{11}) / (M_{01} + M_{10} + M_{11}) = 0 / (2 + 1 + 0) = 0\)
Cosine Similarity

• If \( d_1 \) and \( d_2 \) are two document vectors, then

\[
\cos( d_1, d_2 ) = \frac{(d_1 \cdot d_2)}{||d_1|| \cdot ||d_2||},
\]

where \( \cdot \) indicates vector dot product and \( ||d|| \) is the length of vector \( d \).

• Example:

\[
\begin{align*}
d_1 &= 3 2 0 5 0 0 0 2 0 0 \\
d_2 &= 1 0 0 0 0 0 0 1 0 2
\end{align*}
\]

\[
\begin{align*}
d_1 \cdot d_2 &= 3*1 + 2*0 + 0*0 + 5*0 + 0*0 + 0*0 + 0*0 + 2*1 + 0*0 + 0*2 = 5 \\
||d_1|| &= (3^2+2^2+0^2+5^2+0^2+0^2+0^2+2^2+0^2+0^2+0^2)^{0.5} = (42)^{0.5} = 6.481 \\
||d_2|| &= (1^2+0^2+0^2+0^2+0^2+0^2+0^2+1^2+0^2+2^2)^{0.5} = (6)^{0.5} = 2.245
\end{align*}
\]

\[
\cos( d_1, d_2 ) = .3150, \text{ distance}=1-\cos(d1,d2)
\]
Distance between two values $x$ and $y$ of an attribute $a$ *(Nominal)*

- Value Difference Metric (VDM)- Classes based distance measurements

$$VDM_a(x, y) = \sum_{c=1}^{C} \left| \frac{N_{a,x,c}}{N_{a,x}} - \frac{N_{a,y,c}}{N_{a,y}} \right|$$

The number of output classes

The number of instances in $T$ that have value $x$ for attribute $a$ and output class $c$

Constant, usually 1 or 2

The number of instances in the training set $T$ that have value $x$ for attribute $a$

For example, if an attribute *color* has three values *red*, *green* and *blue*, and the application is to identify whether or not an object is an apple, *red* and *green* would be considered closer than *red* and *blue* because the former two both have similar correlations with the output class *apple*. 
Complex Structure

• For distribution: KL divergence, cross entropy, …
• For trees, graphs: defining graph kernels, …
What we need to know to conduct access log auditing?

• Data Representation
• Data Normalization and Discretization
• Similarity Measurements
• Dimensionality Reduction
Principal Component Analysis (PCA)

- summarization of data with many \( (p) \) variables by a smaller set of \( (k) \) derived (synthetic, composite) variables.
Principal Component Analysis (PCA)

• takes a data matrix of $n$ objects by $p$ variables, which may be correlated, and summarizes it by uncorrelated axes (principal components or principal axes) that are linear combinations of the original $p$ variables

• the first $k$ components display as much as possible of the variation among objects.
Geometric Rationale of PCA

• objects are represented as a cloud of \( n \) points in a multidimensional space with an axis for each of the \( p \) variables

• the centroid of the points is defined by the mean of each variable

• the variance of each variable is the average squared deviation of its \( n \) values around the mean of that variable.

\[
V_i = \frac{1}{n-1} \sum_{m=1}^{n} \left( X_{im} - \overline{X}_i \right)^2
\]
Geometric Rationale of PCA

- degree to which the variables are linearly correlated is represented by their covariance

\[
C_{ij} = \frac{1}{n-1} \sum_{m=1}^{n} (X_{im} - \bar{X}_i)(X_{jm} - \bar{X}_j)
\]
Geometric Rationale of PCA

• The objective of PCA is to rigidly rotate the axes of this p-dimensional space to new positions (principal axes) that have the following properties:
  – ordered such that principal axis 1 has the highest variance, axis 2 has the next highest variance, ..., and axis \( p \) has the lowest variance
  – covariance among each pair of the principal axes is zero (the principal axes are uncorrelated).
2D Example of PCA

- variables $X_1$ and $X_2$ have positive covariance & each has a similar variance.

$\bar{X}_1 = 8.35$  
$\overline{X}_2 = 4.91$

$V_1 = 6.67$  
$V_2 = 6.24$  
$C_{1,2} = 3.42$
**Configuration is Centered**

- each variable is adjusted to a mean of zero (by subtracting the mean from each value).
Principal Components are Computed

- PC 1 has the highest possible variance (9.88)
- PC 2 has a variance of 3.03
- PC 1 and PC 2 have zero covariance.
Auditing: Suspicious or Anomalous?
Auditing: Suspicious or Anomalous?
How Did We Get Here?

- Electronic medical record systems are collaborative systems
- Collaborative systems are about social phenomena
- People *should* form communities
- We should be able to measure deviation from community structure
Collaboration: Users exhibit collaborative behavior in the Vanderbilt StarPanel System

How Did We Learn Collaborative Patterns?

Access logs: EMR Utilization

- A physician requested a lab test: Nov, 9, 2015, 9:00am
- A lab user uploaded a lab test result: Nov, 9, 2015, 4:00pm
- Physician office received the lab test result: Nov, 9, 2015, 4:20pm
- A nurse provided counseling service: Nov, 10, 2015, 10:20am

A Patient’s EHR

A set of Clinical Problems Represented by EMR data
Examples of Accesses

<table>
<thead>
<tr>
<th>patient_pseudonym</th>
<th>enc_identified_num</th>
<th>department_affiliation</th>
<th>Cerner_position</th>
<th>reason</th>
<th>cerner_relationship</th>
<th>user_pseudocln</th>
<th>user_address_dts</th>
<th>hospital_patient_location</th>
</tr>
</thead>
<tbody>
<tr>
<td>442.0000726</td>
<td>1 Obstetrics &amp; Gynecology NMH Physician Office - CPOE</td>
<td>1</td>
<td>Attesting Phys/Prov</td>
<td></td>
<td>OBSTETRICS</td>
<td>47340370</td>
<td>12/10/2010 13:41</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 NMH Resident/Fellow-CPOE</td>
<td>1</td>
<td>Resident- Inpatient Primary Service</td>
<td></td>
<td>OBSTETRICS</td>
<td>47782311</td>
<td>12/10/2010 14:09</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Advanced Practice Clinician - C Nurse Midwife</td>
<td>1</td>
<td></td>
<td></td>
<td>OBSTETRICS</td>
<td>47391708</td>
<td>12/10/2010 14:32</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Patient Care Staff Nurse</td>
<td>1</td>
<td>Covering Staff Nurse</td>
<td></td>
<td>OBSTETRICS</td>
<td>50305400</td>
<td>12/10/2010 14:33</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Patient Care Staff Nurse</td>
<td>1</td>
<td>Primary Staff Nurse</td>
<td></td>
<td>OBSTETRICS</td>
<td>47865070</td>
<td>12/10/2010 14:35</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Maternal &amp; Fetal Medicine NMH Physician-CPOE</td>
<td>1</td>
<td>Patient Care</td>
<td></td>
<td>OBSTETRICS</td>
<td>30450482</td>
<td>12/10/2010 14:52</td>
<td>Labor and Delivery</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 NMH Resident/Fellow-CPOE</td>
<td>1</td>
<td>Resident- Inpatient Covering Service</td>
<td></td>
<td>OBSTETRICS</td>
<td>47923286</td>
<td>12/10/2010 15:00</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 NMH Resident/Fellow-CPOE</td>
<td>1</td>
<td>Resident- Inpatient Consulting Service</td>
<td></td>
<td>OBSTETRICS</td>
<td>50454591</td>
<td>12/10/2010 15:20</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 NMH Resident/Fellow-CPOE</td>
<td>1</td>
<td>Resident- Inpatient Consulting Service</td>
<td></td>
<td>OBSTETRICS</td>
<td>47578593</td>
<td>12/10/2010 15:31</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Anesthesia-CPOE</td>
<td>1</td>
<td>Anesthesiologist</td>
<td></td>
<td>OBSTETRICS</td>
<td>1132285</td>
<td>12/10/2010 15:47</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Patient Care Staff Nurse</td>
<td>1</td>
<td>Covering Staff Nurse</td>
<td></td>
<td>OBSTETRICS</td>
<td>46144460</td>
<td>12/10/2010 16:29</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 NMH Resident/Fellow-CPOE</td>
<td>1</td>
<td>Resident- Inpatient Consulting Service</td>
<td></td>
<td>OBSTETRICS</td>
<td>47578805</td>
<td>12/10/2010 17:20</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 NMH Resident/Fellow-CPOE</td>
<td>1</td>
<td>Resident- Inpatient Primary Service</td>
<td></td>
<td>OBSTETRICS</td>
<td>48531327</td>
<td>12/10/2010 18:56</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Med Student-CPOE</td>
<td>1</td>
<td>Med Student- Inpatient Primary Service</td>
<td></td>
<td>OBSTETRICS</td>
<td>48711960</td>
<td>12/10/2010 19:10</td>
<td>Prentice 13</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Patient Care Staff Nurse</td>
<td>1</td>
<td>Coordinator</td>
<td></td>
<td>OBSTETRICS</td>
<td>40804569</td>
<td>12/10/2010 19:20</td>
<td>Labor and Delivery</td>
</tr>
<tr>
<td>442.0000726</td>
<td>1 Patient Care Staff Nurse</td>
<td>1</td>
<td>Primary Staff Nurse</td>
<td></td>
<td>OBSTETRICS</td>
<td>48126595</td>
<td>12/10/2010 19:46</td>
<td>Prentice 13</td>
</tr>
</tbody>
</table>
Examples of Patient Diagnosis Codes

<table>
<thead>
<tr>
<th>patient_study_id</th>
<th>enc_deiden</th>
<th>DX_codes</th>
</tr>
</thead>
<tbody>
<tr>
<td>442.00000726</td>
<td>1</td>
<td>2165, 65921, 65951, 66401, 66481, V270</td>
</tr>
<tr>
<td>442.0001714</td>
<td>1</td>
<td>V053, V3000</td>
</tr>
<tr>
<td>442.0002396</td>
<td>1</td>
<td>4019, 4111, 41401, 4142, 4739, 49390</td>
</tr>
<tr>
<td>442.0002775</td>
<td>2</td>
<td>1122, 20300, 25000, 27651, 40390, 5845, 5859, 591, 5933, V1005, V1046</td>
</tr>
<tr>
<td>442.0002775</td>
<td>1</td>
<td>1534, 185, 1962, 1974, 20300, 25000, 2809, 40390, 56089, 5849, 5859, 59080, 591, 78791, 7907</td>
</tr>
<tr>
<td>442.0003301</td>
<td>1</td>
<td>76408, 76529, V053, V3100</td>
</tr>
<tr>
<td>442.0004873</td>
<td>1</td>
<td>V270, V8535, 27800, 64911, 64971, 65841, 66401</td>
</tr>
<tr>
<td>442.0005024</td>
<td>1</td>
<td>4019, 72252, V1582</td>
</tr>
<tr>
<td>442.0005968</td>
<td>1</td>
<td>5990, 2724, 511, 4019, 44022</td>
</tr>
<tr>
<td>442.0006352</td>
<td>1</td>
<td>65971, V270</td>
</tr>
<tr>
<td>442.0007008</td>
<td>1</td>
<td>25000, 6144, 99859, V1042</td>
</tr>
<tr>
<td>442.0007371</td>
<td>1</td>
<td>V707, 2859, 33394, 4019, 71690, V420</td>
</tr>
<tr>
<td>442.0007707</td>
<td>1</td>
<td>30000, 49121, 51889, 60000, 7850</td>
</tr>
<tr>
<td>442.0007707</td>
<td>2</td>
<td>78052, V1083, 30001, 496, 0549, 1120, 2768, 30000</td>
</tr>
<tr>
<td>442.0008016</td>
<td>1</td>
<td>V053, V3001</td>
</tr>
<tr>
<td>442.0008405</td>
<td>1</td>
<td>2449, 25080, 4019, 41400, 42731, 4280, 60000, V4581, V5861</td>
</tr>
<tr>
<td>442.0008417</td>
<td>1</td>
<td>2449, 25080, 4019, 41400, 42731, 4280, 60000, V4581, V5861</td>
</tr>
</tbody>
</table>
Community-Based Anomaly Detection System (CADS)


MetaCADS: tripartite graph
CADS: bipartite graph
**MetaCADS:** Concepts of subjects instead of subjects

<table>
<thead>
<tr>
<th></th>
<th>$u_1$</th>
<th>$u_3$</th>
<th>$u_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>$s_1$</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>$s_2$</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>$s_3$</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>$s_4$</strong></td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>$s_6$</strong></td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>$s_7$</strong></td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Access Matrix $A$: cell value 1 indicates a user accessed a patient’s EMR

$$
\begin{array}{ccc}
1 & 1 & 1 \\
1 & 0 & 1 \\
0 & 1 & 1 \\
\end{array}
\times
\begin{array}{ccc}
1 & 1 & 1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 & 1 & 0 \\
\end{array}
= 
\begin{array}{ccc}
\text{4} & \text{2} & \text{2} \\
\text{2} & \text{4} & \text{1} \\
\text{2} & \text{1} & \text{3} \\
\end{array}
$$

Transpose of matrix $A$

Number of subjects $u_1$ and $u_3$ co-accessed

Number of subjects $u_1$ accessed

Number of subjects $u_3$ and $u_6$ co-accessed

Number of subjects $u_3$ and $u_6$ co-accessed
PCA to Model the Community

University Hospital

Vanderbilt Children’s Hospital
Distance measurement of pairs of users

\[ \text{Dis}(u_i, u_j) = \sqrt{\sum_{q=1}^{l} (Z_{qi} - Z_{qj})^2 \times \frac{\lambda_q}{\lambda_{total}}} \]

Projections of \( u_i \) on \( l \) components

Eigen value corresponding to the \( q^{th} \) component

\[ \sum_{i=1}^{l} \frac{\lambda_i}{\sum_{j=1}^{n} \lambda_j} \geq 0.8 \quad \rightarrow \text{Selecting } l \text{ components} \]

\[ \lambda_{total} = \sum_{j=1}^{l} \lambda_j \]
How Do We Set “k”-NN?

• Conductance- a measure of community quality: k=2,3,4?

$$\psi(\beta) = \frac{2}{4}, \psi(\alpha) = \frac{2}{8}, \psi(\gamma) = \frac{2}{\min\{4.12\}}$$

$$\psi(\alpha) < \psi(\beta) = \psi(\gamma)$$

k=3
Minimum conductance at $k=6$
5-Nearest Neighbor Network-Vanderbilt Medical Center
Measuring Deviation from 5-NN

• Every user is assigned a radius $r$:
  – the distance to his fifth nearest neighbor

• Smaller the radius $\rightarrow$ higher density in user’s network

\[
Dev(u_i) = \sqrt{\frac{\sum_{u \in knn_i} (r_j - \bar{r})^2}{k}}
\]

\[
\bar{r} = \frac{\sum_{u \in knn_i} r_j}{k}
\]

\[
\bar{r} = \frac{2+2+2+2+2+3}{6} = 2.17
\]

\[
Dev(q_1) = \frac{(2 - 2.17)^2 \times 5 + (3 - 2.17)^2}{6} = 0.37
\]

Radius for points in the E area are small (e.g., 2), The radius of $q_1$ is 3

Radius for points in the F area are large (e.g., > 6)
CADS on Vanderbilt Access Logs
CADS on Northwestern Access Logs
Example Environments

Electronic Health Records (EHR)

- Vanderbilt University Medical Center “StarPanel” Logs
- 3 months in 2010
- Arbitrary Day
  - ≈ 4,208 users
  - ≈ 1,006 patients
  - ≈ 1,482 diagnoses
  - ≈ 22,014 accesses of subjects
  - ≈ 4,609 assignments of diagnoses
Experimental Design

• Datasets are not annotated for illicit behavior

• We simulated users in several settings to test:
  • Sensitivity to number of patients accessed of a specific users
    • Range from 1 to 120
  • Sensitivity to number of anomalous users
    • simulated users correspond to 0.5% to 5% of total users
    • Number of records accessed fixed to 5
  • Sensitivity to diversity
    • Random number of users (0.5%~5%) and records accessed (1~150)
CADS and Meta CADS dominate other models

MetaCADS dominates when the mix rate (normal and anomalous Users) is low (mix rate = 0.5%), whereas CADS dominates when the mix rate is high (mix rate = 5%)

<table>
<thead>
<tr>
<th>MODEL</th>
<th>0.5%</th>
<th>2%</th>
<th>5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetaCADS</td>
<td>0.92±0.02</td>
<td>0.90±0.01</td>
<td>0.87±0.03</td>
</tr>
<tr>
<td>CADS</td>
<td>0.91±0.01</td>
<td>0.94±0.02</td>
<td>0.94±0.01</td>
</tr>
<tr>
<td>KNN</td>
<td>0.75±0.02</td>
<td>0.73±0.03</td>
<td>0.72±0.04</td>
</tr>
<tr>
<td>PCA</td>
<td>0.72±0.03</td>
<td>0.74±0.02</td>
<td>0.75±0.03</td>
</tr>
<tr>
<td>HVU</td>
<td>0.68±0.03</td>
<td>0.68±0.03</td>
<td>0.68±0.03</td>
</tr>
</tbody>
</table>

Because when the number of anomalous users increases, MetaCADS enable anomalous users to form communities, thus lowering their deviation scores.
MetaCADS and CADS figure out which user is anomalous, but do not figure out whom the user illegally access?
For subject $s_3$, five users accessed its record, is there any anomalous (illegal) access here?

Who Illegally access $s_3$?


User Relationship Measurement

<table>
<thead>
<tr>
<th></th>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$u_3$</th>
<th>$u_4$</th>
<th>$u_5$</th>
<th>$u_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$s_2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_3$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$s_4$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_5$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$s_6$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$s_7$</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$Sim(u_i, u_j) = \frac{U_i \cdot U_j}{||U_i|| \times ||U_j||}$

<table>
<thead>
<tr>
<th></th>
<th>$u_1$</th>
<th>$u_2$</th>
<th>$u_4$</th>
<th>$u_5$</th>
<th>$u_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1$</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_2$</td>
<td>0.50</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_4$</td>
<td>0.50</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$u_5$</td>
<td>0.50</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>$u_6$</td>
<td>0.58</td>
<td>0.29</td>
<td>0.29</td>
<td>0.29</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Changes of network similarity with the removal of an access

$u_6$ illegally access $s_3$?

Remove Access: $u_6 \rightarrow s_3$

<table>
<thead>
<tr>
<th>Network</th>
<th>Similarity</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>$u_1, u_2, u_4, u_5, u_6$</td>
<td>0.59</td>
<td>5</td>
</tr>
<tr>
<td>$u_1, u_2, u_4, u_5$</td>
<td>0.75</td>
<td>4</td>
</tr>
</tbody>
</table>

Removal access | Changes of network similarity
-----------------|----------------------------------
$u_6 \rightarrow s_3$  | 0.16

Similarity Increase

Access score: the score for access $u_6 \rightarrow s_3$
Scores of one week accesses generated from an EHR system

Accesses with high scores are considered as anomalous ones, which need to be investigated.
Scores of one week edits generated from wiki
Evaluation

For a random user, verifying how number of simulated access injected into this user influence the performances of SNAD

For a fixed number of simulated accesses, verifying how number of intruded users influence the performances of SNAD

The number of simulated accesses and intruded users are both diverse
Model Evaluation-setting 1

For a random user, injecting simulated accesses

<table>
<thead>
<tr>
<th></th>
<th>s_1</th>
<th>s_2</th>
<th>s_3</th>
<th>...</th>
<th>s_i</th>
<th>...</th>
<th>s_n</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
<td>...</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td></td>
<td>...</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>...</td>
<td>0</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
</tbody>
</table>

For a random user, injecting simulated accesses
(a) EHR

(b) Wiki
Model Evaluation-setting 2

Fixing number of simulated accesses, number of intruders is random

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>$s_2$</td>
<td>$s_3$</td>
<td>...</td>
<td>$s_i$</td>
<td>...</td>
<td>$s_n$</td>
</tr>
</tbody>
</table>

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>0</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Intruder_1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>Intruder_2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>Intruder_k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
(a) EHR

(b) Wiki
Model Evaluation-setting 3

Fixing number of intruders, number of access is random

\[
\begin{array}{cccccccc}
  s_1 & s_2 & s_3 & \ldots & s_i & \ldots & s_n \\
\end{array}
\]

1 1 0 \ldots 0 \ldots 0 \quad \text{Intruder}_1

0 1 1 \ldots 1 \ldots 1 \quad \text{Intruder}_2

\ldots

0 1 1 \ldots 1 \ldots 0 \quad \text{Intruder}_k
Conclusions

It is an effective way by using social network analysis to detect anomalous usages of electronic health records, such as CADS and SNAD