Finding Outstanding Aspects and Contrast Subspaces

Jian Pei
School of Computing Science
Simon Fraser University
jpe@cs.sfu.ca
CHIRC

• Computational Health Intelligence Research Centre
  – Population health powered by big data
  – Healthcare business intelligence
  – Predictive health analytics

• A collaborative research initiative with industry leaders

• Technology transferred to industry
  – Multi-million US dollars financial gain per year for industry partners
In what aspect is he most similar to cases of **coronary artery disease** and, at the same time, dissimilar to **adiposity**?

**Symptoms:**
overweight,
high blood pressure,
back pain,
short of breadth,
chest pain,
cold sweat
…
Fraud Suspect Analysis

• An insurance analyst is investigating a suspicious claim
• How is the claim compared with the normal and fraud claims?
  – In what aspects the suspicious case is most similar to fraudulent cases and different from normal claims?
Don’t You Ever Google Yourself?

• Big data makes one know oneself better
• 57% American adults search themselves on Internet
  – Good news: those people are better paid than those who haven’t done so! (Investors.com)
• Egocentric analysis becomes more and more important with big data
Egocentric Analysis

- How am I different from (more often than not, better than) others?
- In what aspects am I good?
Contrast Subspace Finding

• Given a set of labeled objects in two classes
• For a query object q that is also labeled, the contrast subspace is the one where q is most likely to belong to the target class against the other class
Related Work

• Finding patterns and models that manifest drastic differences from one class against the other
  – Example: emerging patterns
• Subspace outlier detection
  – The query object may not be an outlier
• Typicality queries do not consider subspaces
Problem Formulation

• Find subspaces maximizing $LC_S(q) = \frac{L_S(q \mid O_+)}{L_S(q \mid O_-)}$

• To avoid triviality, consider only subspaces where $L_S(q \mid O_+) \geq \delta$
Density Estimation

- Density estimated by
  \[ L_S(q \mid O) = \hat{f}_S(q, O) = \frac{1}{|O|\sqrt{2\pi} h_S} \sum_{o \in O} e^{-\frac{\text{dist}_S(q,o)^2}{2h^2_S}} \]
  
- Then,
  \[ L_{C_{S}}(q, O_+, O_-) = \frac{\hat{f}_S(q, O_+)}{\hat{f}_S(q, O_-)} = \frac{|O_-| h_{S_{-}}}{|O_+| h_{S_{+}}} \cdot \sum_{o \in O_+} e^{-\frac{\text{dist}_S(q,o)^2}{2h^2_{S_{+}}}} \]
  \[ \sum_{o \in O_{-}} e^{-\frac{\text{dist}_S(q,o)^2}{2h^2_{S_{-}}}} \]
Complexity

• MAX SNP-hard
  – Reduction from the emerging pattern mining problem
• Impossible to design a good approximation algorithm
A Monotonic Bound

- $L_S(q \mid O_+)$ is not monotonic in subspaces
- Develop an upper bound of $L_S(q \mid O_+)$, which is monotonic in subspaces
  - Sort all the dimensions in their standard deviation descending order
  - Let $\mathcal{S}$ be the set of children of $S$ in the subspace set enumeration tree using the standard deviation descending order

$$L^*_S(q \mid O_+) = \frac{1}{|O_+| \sqrt{2 \pi \sigma'_{\min} h'_{\text{opt-min}}} \sum_{o \in O_+} e^{-\frac{-\text{dist}_S(q,o)^2}{2\sigma'_{\min} h'_{\text{opt-min}}}}} \sum_{o \in O_+} e^{-\frac{-\text{dist}_S(q,o)^2}{2\sigma'_{\min} h'_{\text{opt-min}}}}$$

- $\sigma'_{\min} = \min\{\sigma_{S'} \mid S' \in \mathcal{S}\}$, $h'_{\text{opt-min}} = \min\{h_{S'_{\text{opt}}} \mid S' \in \mathcal{S}\}$, and $h'_{\text{opt-max}} = \max\{h_{S'_{\text{opt}}} \mid S' \in \mathcal{S}\}$
Monotonic Bound

For a query object $q$, a set of objects $O$, and subspaces $S_1$, $S_2$ such that $S_1$ is an ancestor of $S_2$ in the subspace set enumeration tree using the standard deviation descending order in $O_+$, $L_{S_1}^*(q \mid O_+) \geq L_{S_2}(q \mid O_+)$.

Baseline algorithm time complexity:

$$O(2^{|D|} \cdot (|O_+| + |O_-|))$$
Bounding Using Neighborhoods

- Divide the neighborhood of an object into two parts $N^\epsilon_S(q) = \{ o \in O \mid dist_S(q, o) \leq \epsilon \}$ and the rest.

- Then, $L_S(q \mid O) = L_{N^\epsilon_S}(q \mid O) + L_{rest}^S(q \mid O)$

  $$L_{N^\epsilon_S}(q \mid O) = \frac{1}{|O| \sqrt{2\pi h_S}} \sum_{o \in N^\epsilon_S(q)} e^{-\frac{dist_S(q, o)^2}{2h_S^2}}$$

  $$L_{rest}^S(q \mid O) = \frac{1}{|O| \sqrt{2\pi h_S}} \sum_{o \in O \setminus N^\epsilon_S(q)} e^{-\frac{dist_S(q, o)^2}{2h_S^2}}$$
Bounding the Rest

• Let $\overline{\text{dist}}_S(q \mid O)$ be the maximum distance between $q$ and all objects in $O$ in subspace $S$

$$\frac{|O| - |N^c_S(q)|}{|O|\sqrt{2\pi}h_S} \cdot e^{\frac{\overline{\text{dist}}_S(q,O)^2}{2h^2_S}} \leq L^\text{rest}_S(q \mid O) \leq \frac{|O| - |N^c_S(q)|}{|O|\sqrt{2\pi}h_S} \cdot e^{-\frac{\epsilon^2}{2h^2_S}}$$
For a query object $q$, a set of objects $O$ and $\epsilon \geq 0$,

$$LL^\epsilon_S(q \mid O) \leq L_S(q \mid O) \leq UL^\epsilon_S(q \mid O)$$

where

$$LL^\epsilon_S(q \mid O) = \frac{1}{|O|\sqrt{2\pi h_S}} \left( \sum_{o \in N^\epsilon_S(q)} e^{-\frac{\text{dist}^\epsilon_S(q,o)^2}{2h^2_S}} + (|O| - |N^\epsilon_S(q)|)e^{-\frac{\text{dist}^\epsilon_S(q,O)^2}{2h^2_S}} \right)$$

and

$$UL^\epsilon_S(q \mid O) = \frac{1}{|O|\sqrt{2\pi h_S}} \left( \sum_{o \in N^\epsilon_S(q)} e^{-\frac{\text{dist}^\epsilon_S(q,o)^2}{2h^2_S}} + (|O| - |N^\epsilon_S(q)|)e^{-\frac{\epsilon^2}{2h^2_S}} \right)$$

For a query object $q$, a set of objects $O_+$, a set of objects $O_-$, and $\epsilon \geq 0$,

$$LC_S(q) \leq \frac{UL^\epsilon_S(q \mid O_+)}{LL^\epsilon_S(q \mid O_-)}.$$
Algorithm

Algorithm 1 $CSMiner(q, O_+, O_-, \delta, k)$

**Input:** $q$: a query object, $O_+$: the set of objects belonging to $C_+$, $O_-$: the set of objects belonging to $C_-$, $\delta$: a likelihood threshold, $k$: positive integer

**Output:** $k$ subspaces with the highest likelihood contrast

1. let $Ans$ be the current top-$k$ list of subspaces, initialize $Ans$ as $k$ null subspaces associated with likelihood contrast 0
2. for each subspace $S$ in the subspace set enumeration tree, searched in the depth-first manner do
3. if $UL_S^c(q \mid O_+) \geq \delta$ and $\exists S' \in Ans$ s.t. $UL_S^c(q \mid O_+) \geq UL_{S'}^c(q \mid O_-)$ then
4. calculate $L_S(q \mid O_+), L_S(q \mid O_-)$ and $LC_S(q)$; // refining
5. if $L_S(q \mid O_+) \geq \delta$ and $\exists S' \in Ans$ s.t. $LC_S(q) > LC_{S'}(q)$ then
6. insert $S$ into the top-$k$ list
7. end if
8. end if
9. if $L_S^*(q \mid O_+) < \delta$ then
10. prune all super-spaces of $S$;
11. end if
12. end for
13. return $Ans$;
Dimensionality of Inlying Contrast Subspaces

(a) BCW

(b) CMSC

(c) Glass

(d) PID

(e) Waveform

(f) Wine
Dimensionality of Outlying Contrast Subspaces

(a) BCW

(b) CMSC

(c) Glass

(d) PID

(e) Waveform

(f) Wine
Runtime

(a) w.r.t $\delta$ ($k = 10, r = 0.4$)  (b) w.r.t data set size ($k = 10, \delta = 0.01, r = 0.4$)  (c) w.r.t dimensionality ($k = 10, \delta = 0.01, r = 0.4$)
In Which Aspects Johnson Is Good?

J. Pei: Finding Outstanding Aspects and Contrast Subspaces
Fraud Investigation

- Given a set of claims in an insurance company
- For a claim $c$, in which aspects $c$ is most different from the other claims?
Outlying/Outstanding Aspect Mining

• Given a set of objects in a multi-dimensional space
• For an object q, find the subspaces where q is most unusual compared to the rest of the data
Differences from Outlier Detection

- Outlier detection finds objects that are different from the rest of the data.
- The query object in outlying aspect finding may not be an outlier.
Problem Formulation

• A set of objects $O$ in full space
  \[ D = \{D_1, \ldots, D_d\} \]

• Query object $q$

• The density of $q$ measures how outlying (uncommon) $q$ is
  – Density estimation
    \[ \hat{f}_h(o) = \frac{1}{n} \sum_{i=1}^{n} K_h(o - o_i) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{o - o_i}{h} \right) \]

• Find a subspace where the density of $q$ is lowest?
Why Rank Statistics?

- Densities in different subspaces are not comparable
- We compare the same set of objects in different subspaces
- Rank statistics

\[ rank_S(o) = |\{o' | o' \in O, OutDeg(o') < OutDeg(o)\}| + 1 \]
Unsupervised Problem Formulation

Given a set of objects $O$ in a multidimensional space $D$, a query object $q \in O$ and a maximum dimensionality threshold $0 < \ell \leq |D|$, a subspace $S \subseteq D$ ($0 < |S| \leq \ell$) is called a minimal outlying subspace of $q$ if

1. (Rank minimality) there does not exist another subspace $S' \subseteq D$ ($S' \neq \emptyset$), such that $\text{rank}_{S'}(q) < \text{rank}_S(q)$; and

2. (Subspace minimality) there does not exist another subspace $S'' \subset S$ such that $\text{rank}_{S''}(q) = \text{rank}_S(q)$.

The problem of outlying aspect mining is to find the minimal outlying subspaces of $q$. 

J. Pei: Finding Outstanding Aspects and Contrast Subspaces
Density Estimation for Ranking

\[ \hat{f}_S(q) \sim \tilde{f}_S(q) = \sum_{o \in O} e^{-\sum_{D_i \in S} \frac{(q.D_i - o.D_i)^2}{2h^2_{D_i}}} \]

**Invariance**

Given a set of objects \( O \) in space \( S = \{D_1, \ldots, D_d\} \), define a linear transformation \( g(o) = (a_1 o.D_1 + b_1, \ldots, a_d o.D_d + b_d) \) for any \( o \in O \), where \( a_1, \ldots, a_d \) and \( b_1, \ldots, b_d \) are real numbers. Let \( O' = \{g(o)|o \in O\} \) be the transformed data set. For any objects \( o_1, o_2 \in O \) such that \( \hat{f}_S(o_1) > \hat{f}_S(o_2) \) in \( O \), \( \tilde{f}_S(g(o_1)) > \tilde{f}_S(g(o_2)) \) if the product kernel is used and the bandwidths are set using Härdle’s rule of thumb.
Algorithm Framework

Algorithm 2 The framework of OAMiner

Input: a set of objects \( O \) and query object \( q \in O \)

Output: the set of minimal outlying subspaces for \( q \)

1: initialize \( r_{best} \leftarrow |O| \) and \( Ans \leftarrow \emptyset \);
2: remove \( D_i \) from \( D \) if the values of all objects in \( D_i \) are identical;
3: compute \( rank_{D_i}(q) \) in each dimension \( D_i \in D \);
4: sort all dimensions in \( rank_{D_i}(q) \) ascending order;
5: for each subspace \( S \) searched by traversing the set enumeration tree in a depth-first manner do
6: compute \( rank_S(q) \);
7: if \( rank_S(q) < r_{best} \) then
8: \( r_{best} \leftarrow rank_S(q), Ans \leftarrow \{S\} \);
9: end if
10: if \( rank_S(q) = r_{best} \) and \( S \) is minimal then
11: \( Ans \leftarrow Ans \cup \{S\} \);
12: end if
13: if a subspace pruning condition is true then
14: prune all super-spaces of \( S \)
15: end if
16: end for
17: return \( Ans \)
Pruning Rule 1

• If \( \text{rank}_S(q) = 1 \), according to the dimensionality minimality condition in the problem definition, all super-spaces of S can be pruned.

• Pruning on other ranks or density values?
  – Neither rank nor density is not monotonic with respect to subspaces
Reducing Density Estimation Cost

- To obtain the exact rank statistics in a subspace, the query object has to compare with every other object.
- By estimating density values using neighborhood, density computation can be reduced.
Cross Subspace Pruning

• For subspaces $S \subset S'$, by estimating the bounds of possible changes in density, then the range of the rank in $S'$ can be estimated by the rank in $S$

• Some subspaces can be pruned using the ranges
Distribution of Ranks

(a) Guards
(b) Forwards
(c) Centers
(d) Breast cancer
(e) Climate model
(f) Concrete slump
(g) Parkinsons
(h) Wine
Distribution of # Outlying Aspects

(a) Guards
(b) Forwards
(c) Centers
(d) Breast cancer
(e) Climate model
(f) Concrete slumph
(g) Parkinsons
(h) Wine
Computational Performance

![Graphs showing computational performance](image)

J. Pei: Finding Outstanding Aspects and Contrast Subspaces

37
Conclusions

• Finding outlying/outstanding aspects and contrast subspaces has many applications
• Computationally, it is challenging – even cannot be approximated well
• Future work
  – Faster algorithms
  – More effective measures
  – Scaling out
Papers
