Support Vector Subset Scan for Spatial Pattern Detection

Dylan Fitzpatrick, Yun Ni, and Daniel B. Neill
Event and Pattern Detection Laboratory
Carnegie Mellon University

This work was partially supported by NSF grant IIS-0953330
Detecting Spatial Clusters

Given a set of data streams, can we find regions with counts significantly higher than expected?

**Goal:** Method with high detection power that is computationally efficient

**Problem:** Regions may be highly irregular in shape. $2^N$ different subsets.
Detecting Spatial Clusters

Spatial Scan Statistic (Kulldorff, 1997):

- Searches over circular regions

  **High detection power** for affected regions of corresponding shape

  **Low detection power** for irregular clusters
Detecting Irregular Spatial Clusters

Fast Subset Scan (Neill, 2011):

Finds most anomalous subset over entire region (or constrained subregions) efficiently and exactly.

Can we impose spatial constraints without losing detection power for subtle and irregular patterns?
Detecting Irregular Spatial Clusters

Fast Subset Scan (Neill, 2011):

Finds most anomalous subset over entire region (or constrained subregions) efficiently and exactly

Can we impose spatial constraints without losing detection power for subtle and irregular patterns?
Expectation-Based Scan Statistics

Poisson Example:

\[ H_0 : c_i \sim \text{Poisson}(b_i) \]
\[ H_1 : c_i \sim \text{Poisson}(qb_i), q > 1 \]

\[ F(S) = \max_{q>1} \log \frac{P(\text{Data}|H_1(S))}{P(\text{Data}|H_0)} \]

VS.

Large subset, moderate risk

VS.

Small pattern, high risk
For a data set $D$, score function $F(S)$ satisfies the Additive Linear Subset Scanning (ALTSS) property if for all $S \subseteq D$,

$$F(S) = \max_{q > 1} F(S|q) \text{ where } F(S|q) = \sum_{s_i \in S} \lambda_i$$

and where $\lambda_i$ depends only on observed count $c_i$, expected count $b_i$, and fixed relative risk $q$. 

Adding Element-Specific Penalties

Penalized Fast Subset Scan (Speakman et al., 2015):
Adding Element-Specific Penalties

Penalized Fast Subset Scan (Speakman et al., 2015):

<table>
<thead>
<tr>
<th>Distribution</th>
<th>$\lambda_i(q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson</td>
<td>$x_i (\log q) + \mu_i (1 - q)$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$x_i \frac{\mu_i}{\sigma_i^2} (q - 1) + \mu_i \frac{\mu_i}{\sigma_i^2} (\frac{1-q^2}{2})$</td>
</tr>
<tr>
<td>exponential</td>
<td>$x_i \frac{1}{\mu_i} (1 - \frac{1}{q}) + \mu_i \frac{1}{\mu_i} (-\log q)$</td>
</tr>
<tr>
<td>binomial($p_0$)</td>
<td>$x_i \log(q \frac{1-p_0}{1-qp_0}) + \log(\frac{1-qp_0}{1-p_0})$</td>
</tr>
</tbody>
</table>
Adding Element-Specific Penalties

Penalized Fast Subset Scan (Speakman et al., 2015):

Element-specific terms can be added to score function while maintaining additive property

\[ F_{\text{penalized}}(S') = \max_{q > 1} \sum_{s_i \in S} (\lambda_i + \Delta_i) \]

**Easy to interpret:** \( \Delta_i \) terms are the prior log-odds of data point \( s_i \) being in the true affected subset.

**Easy to maximize:** For fixed relative risk \( q \), only include points with positive overall contribution. Optimal subset can be found by considering \( O(N) \) values of \( q \).
Support Vector Machine

Classification algorithm that finds the separating hyperplane which maximizes the margin between positive and negative data points

Support Vector Machine

\[
\min_{\xi, w, b} \frac{1}{2}||w||^2 + C \sum_{i=1}^{N} \xi_i
\]

\[
\xi_i \geq 0, \forall i = 1, \ldots, N
\]

\[
y_i(w \cdot \phi(x_i) - b) \geq 1 - \xi_i, \forall i = 1, \ldots, N
\]

where:

- weight vector \(w\) and bias term \(b\) define a hyperplane
- \(\xi_i\) terms allow for approximation in case data are not linearly separable
- \(\phi\) is a transformation to high-dimensional feature space allowing for non-linear decision boundaries
- \(w \cdot \phi(x_i) - b\) is a measure of distance from point \(x_i\) to the hyperplane
Support Vector Subset Scan (SVSS)

**Intuition:** Find anomalous subset with large margin between affected and unaffected points

**Result:** Irregular but spatially coherent regions
Let \( x_i \) be the spatial coordinates of point \( s_i \), let \( \alpha_i \in \{0, 1\} \) indicate presence/absence of point \( i \) in \( S \), and let \( y_i = 2\alpha_i - 1 \)

\[
\min_{\alpha, \xi, w, b} \frac{1}{2} \|w\|^2 + C_0 \sum_{i=1}^{N} \xi_i - C_1 F(\alpha)
\]

\[
\alpha_i \in \{0, 1\}, \forall i = 1, \ldots, N
\]

\[
\xi_i \geq 0, \forall i = 1, \ldots, N
\]

\[
(2\alpha_i - 1)(w \cdot \phi(x_i) - b) \geq 1 - \xi_i, \forall i = 1, \ldots, N
\]
SVSS Objective Function

Equivalently,

$$
\min_{\alpha, \xi, w, b} \frac{1}{2} \|w\|^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\alpha)
$$

$$
\alpha_i \in \{0, 1\}, \forall i = 1, ..., N
$$

$$
\xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(w \cdot \phi(x_i) - b))
$$
SVSS Objective Function

Equivalently,

\[
\min_{\alpha, \xi, w, b} \frac{1}{2} \| w \|^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\alpha)
\]

\[\alpha_i \in \{0, 1\}, \forall i = 1, ..., N\]

\[\xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(w \cdot \phi(x_i) - b))\]

**Problem:** Objective is not convex. We optimize with alternate minimization and multiple random restarts.
SVSS Objective Function

Equivalently,

$$\min_{\alpha, \xi, w, b} \frac{1}{2} \|w\|^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\alpha)$$

$$\alpha_i \in \{0, 1\}, \forall i = 1, \ldots, N$$

$$\xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(w \cdot \phi(x_i) - b))$$

PFSS Problem

Element-specific penalties = Distance to SVM hyperplane
SVSS Objective Function

Equivalently,

\[
\min_{\alpha, \xi, \mathbf{w}, b} \left( \frac{1}{2} \lVert \mathbf{w} \rVert^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) \right) - C_1 F(\alpha)
\]

\[\alpha_i \in \{0, 1\}, \forall i = 1, ..., N\]

\[\xi_i(\alpha_i) = \max(0, 1 - (2\alpha_i - 1)(\mathbf{w} \cdot \phi(x_i) - b))\]

SVM Problem

Binary data labels = Included/Not included in subset
SVSS Algorithm

Algorithm 1 Support Vector Subset Scan

procedure SVSS(c, b, x, C₀, C₁) ▷ Counts c, expectations b, and coordinates x

ξᵢ(αᵢ) ← 0, ∀i = 1, ..., N

while The optimal subset is changing do

\[
\begin{align*}
\max_{α} & \quad F(α) - \frac{C₀}{C₁} ∑_{i=1}^{N} ξᵢ(αᵢ) \\
\min_{ξ, w, b} & \quad \frac{1}{2}||w||^2 + C₀ ∑_{i=1}^{N} ξᵢ(αᵢ)
\end{align*}
\]

▷ Fix w, b and optimize over α

▷ Fix α, and optimize over w, b

end while

return α

end procedure
SVSS Algorithm

Algorithm 1 Support Vector Subset Scan

procedure SVSS(c, b, x, C₀, C₁)

\[ \xi_i(\alpha_i) \leftarrow 0, \forall i = 1, ..., N \]

\[ \text{while The optimal subset is changing do} \]

\[ \max_{\alpha} F(\alpha) - C₀/C₁ \sum_{i=1}^{N} \xi_i(\alpha_i) \]

\[ \min_{\xi, w, b} \frac{1}{2} ||w||^2 + C₀ \sum_{i=1}^{N} \xi_i(\alpha_i) \]

\[ \text{end while} \]

return \( \alpha \)

end procedure

PFSS

SVM
SVSS Algorithm

Algorithm 2 Support Vector Subset Scan (random restarts)

procedure SVSS(c, b, x, T_{max}, C_0, C_1) \triangleright Counts c, expectations b, and coordinates x
\[
\text{min}\_\text{score} \leftarrow \infty
\]
for \( t := 1 \) to \( T_{max} \) do \triangleright \( T_{max} \) random restarts
\[
\xi_i(\alpha_i) \leftarrow \text{Uniform}(-C_0, C_0), \forall i = 1, \ldots, N
\]
while The optimal subset is changing do
\[
\begin{align*}
\max_{\alpha} F(\alpha) &- C_0/C_1 \sum_{i=1}^{N} \xi_i(\alpha_i) \triangleright \text{Fix w, b and optimize over } \alpha \\
\min_{\xi, w, b} \frac{1}{2}\|w\|^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) &\quad \triangleright \text{Fix } \alpha, \text{ and optimize over } w, b
\end{align*}
\]
end while
\[
\text{score} \leftarrow \frac{1}{2}\|w\|^2 + C_0 \sum_{i=1}^{N} \xi_i(\alpha_i) - C_1 F(\alpha)
\]
if \( \text{score} < \text{min}\_\text{score} \) then
\[
\begin{align*}
\text{min}\_\text{score} &\leftarrow \text{score} \\
\alpha_{min} &\leftarrow \alpha
\end{align*}
\]
end if
end for
return \( \alpha_{min} \)
end procedure
Computing Penalties

$$\arg\max_\alpha F(\alpha) - \frac{C_0}{C_1} \sum_{i=1}^{N} \xi_i(\alpha_i)$$

$$\xi_i(\alpha_i) = \begin{cases} 
\max(0, 1 - w \cdot \phi(x_i) + b), & y_i = 2\alpha_i - 1 = +1 \\
\max(0, 1 + w \cdot \phi(x_i) - b), & y_i = 2\alpha_i - 1 = -1 
\end{cases}$$

How to fit into PFSS framework?

**Needed:** Element-specific penalties for included sites
Computing Penalties

EQUIVALENT:

$$\arg\max_{\alpha} F(\alpha) - \frac{C_0}{C_1} \sum_{i=1}^{N} \alpha_i \Delta_i$$

$$\Delta_i = \max(0, 1 - w \cdot \phi(x_i) + b) - \max(0, 1 + w \cdot \phi(x_i) - b)$$

$$= \begin{cases} 
  w \cdot \phi(x_i) - b + 1, & w \cdot \phi(x_i) - b \geq 1 \\
  2(w \cdot \phi(x_i) - b), & w \cdot \phi(x_i) - b \in (-1, 1) \\
  w \cdot \phi(x_i) - b - 1, & w \cdot \phi(x_i) - b \leq -1 
\end{cases}$$

$$= \left[ w \cdot \phi(x_i) - b > -1 \right] (w \cdot \phi(x_i) - b + 1) +$$

$$\left[ w \cdot \phi(x_i) - b < 1 \right] (w \cdot \phi(x_i) - b - 1)$$
Improvement Over Iterations

Expectation = 100 for all sites

Unaffected points ~ Poisson(100)

Affected points ~ Poisson(120)
Improvement Over Iterations
Improvement Over Iterations
Improvement Over Iterations
Ranking Disconnected Regions

How can we rank the connected regions of the best subset?

**Solution**: Maximize penalized log-likelihood ratio over connected components of SVM decision boundary
Tuning model parameters

Goal: Find parameter combination that generates best subset with high log-likelihood ratio (LLR) and some minimum level of geometric compactness
Tuning model parameters

Tuning procedure:
1. Define measure of geometric compactness \( K \) (Duzcmal et al., 2006):
   \[
   K(z) = \frac{4\pi A(z)}{H(z)^2}
   \]
   where \( A(z) = \text{Area of } z \), \( H(z) = \text{Perimeter of convex hull of } z \)

2. Maximize LLR of best subset over parameter settings with top SVM component meeting minimum compactness threshold

VS.
Detecting Letter-Shaped Regions

\[ c_i \sim \text{Poisson}(100) \]
\[ c_i \sim \text{Poisson}(120) \]
\[ c_i \sim \text{Poisson}(140) \]
\[ c_i \sim \text{Poisson}(160) \]
\[ c_i \sim \text{Poisson}(180) \]

All points: \( b_i = 100 \)
Detecting Letter-Shaped Regions

Best connected SVM region
Detecting Letter-Shaped Regions

2nd Best connected SVM region
Detecting Letter-Shaped Regions
Detecting Letter-Shaped Regions

4th Best connected SVM region
Evaluation Framework

• 2000 observations generated from Poisson distribution

• Generated random, irregular-shaped regions of varying length with elevated counts
  – Unaffected points: \( c_i \sim \text{Poisson}(100) \)
  – Affected points: \( c_i \sim \text{Poisson}(115) \)
  – \( b_i = 100 \) for all points

• Compared precision and recall of top pattern at each length against:
  – Fast subset Scan (Neill, 2011)
  – Circular scan statistic (Kulldorff, 1997)
  – Upper level set scan statistic (Patil and Taillie, 2007)
Detecting Pothole Hotspots

Data:
• Pothole reports at city block level from City of Pittsburgh 311 system

Timeframe:
• Expected counts estimated from 2008-2011 control period
• Actual counts generated from 2012-2013

Can we identify roads or neighborhoods in need of maintenance?
# Top 5 Pothole Hotspots

<table>
<thead>
<tr>
<th>Rank</th>
<th># of Points</th>
<th>Relative Risk (MLE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>17</td>
<td>3.2</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>3.0</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>2.8</td>
</tr>
<tr>
<td>4</td>
<td>12</td>
<td>3.9</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>2.3</td>
</tr>
</tbody>
</table>

*Pattern shown to right*
Conclusion

Support Vector Subset Scan (SVSS) is a new method for detecting localized and irregularly shaped patterns which are spatially separated from non-anomalous data.

In simulated experiments, SVSS showed high precision and recall on the task of detecting irregularly shaped patterns relative to competing methods.

We demonstrated the real-world utility of SVSS by applying it to pothole hotspot detection in Pittsburgh roadways.
Thank you

djfitzpa@cmu.edu
neill@cs.cmu.edu