Predicting Building Contamination Using Machine Learning

Shawn Martin and Sean McKenna

Sandia National Laboratories
Albuquerque, NM

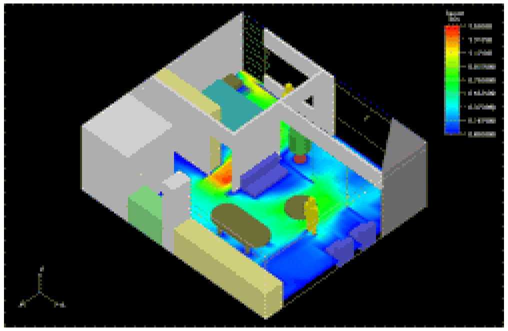
12/13/2007

Why Model Building Contamination?

- In the event of disaster …
  - Should building be evacuated or should residents shelter in place?
  - Should ducts be closed or purged?
  - Where is contamination, and where is it going?
- After the disaster …
  - Where should measurements be taken?
  - Where is residual contamination?
  - What is the best way to clean up the building?
- Before the next disaster …
  - Models can be used to design new buildings to minimize future events.

Current Building Models

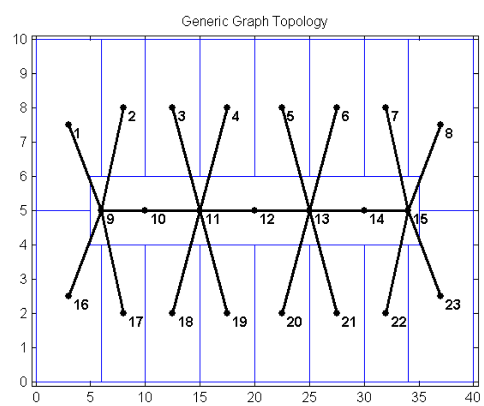
- Models are used to predict airflow throughout a building.
  - Predict Heating, Ventilation, and Air Conditioning (HVAC) operation.
  - Predict how smoke would travel through a building.
  - Predict how biological or chemical contaminants would travel in an attack.
- Computational Fluid Dynamics (CFD)
  - Very precise, but computationally intensive.
  - Can be used for single rooms or small buildings.

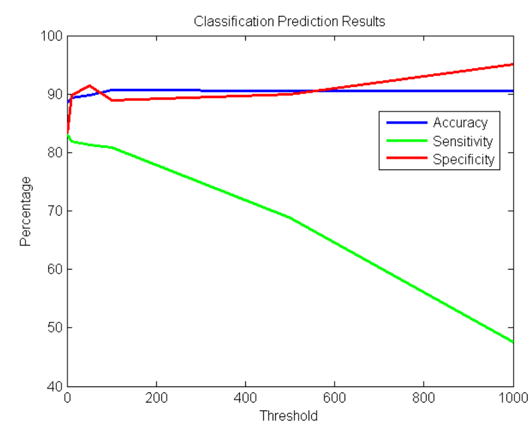


- Multizonal Methods
  - Models air flow between rooms with well-mixed air.
  - Widely used, best current compromise between accuracy and speed.
- Statistical Methods
  - Kriging, Kalman Filtering, Bayesian Monte Carlo.

Machine Learning Building Model

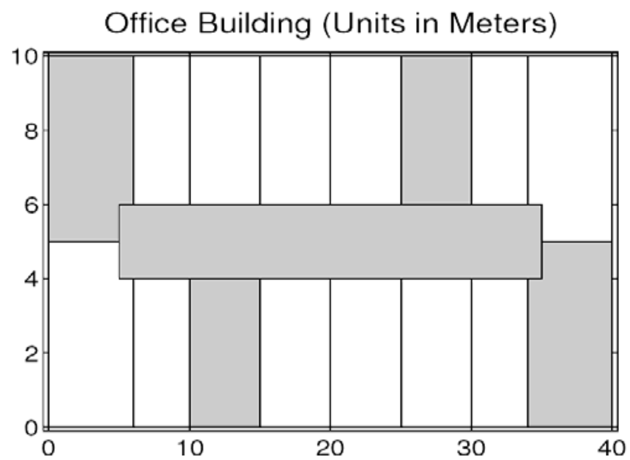
- Proceeds in two steps:
  - Train Support Vector Machine (SVM) using multiple contamination events.
  - Use SVM model to predict results of a given event.
- Advantages:
  - Most of the computational effort is in training the model.
  - *Predictions can be made in real-time.*
- Disadvantages:
  - Loss of accuracy compared to CFD-type models.
  - Large training sets required.
- Similar to statistical methods, especially Bayesian Monte Carlo approach.

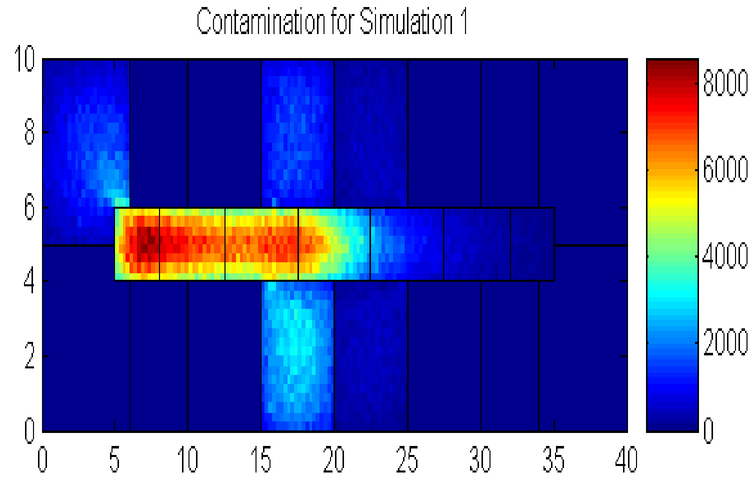




Building Simulation Data

- Due to lack of real world data, we generated simulations of a simple 2-D office building using particle transport model.
- We generated two datasets
  - Dataset A: 120 simulations with randomly chosen configurations of the building (open/closed doors, advection, diffusion) but same source location.
  - Dataset B: 250 simulations with randomly chosen configurations with different source locations.

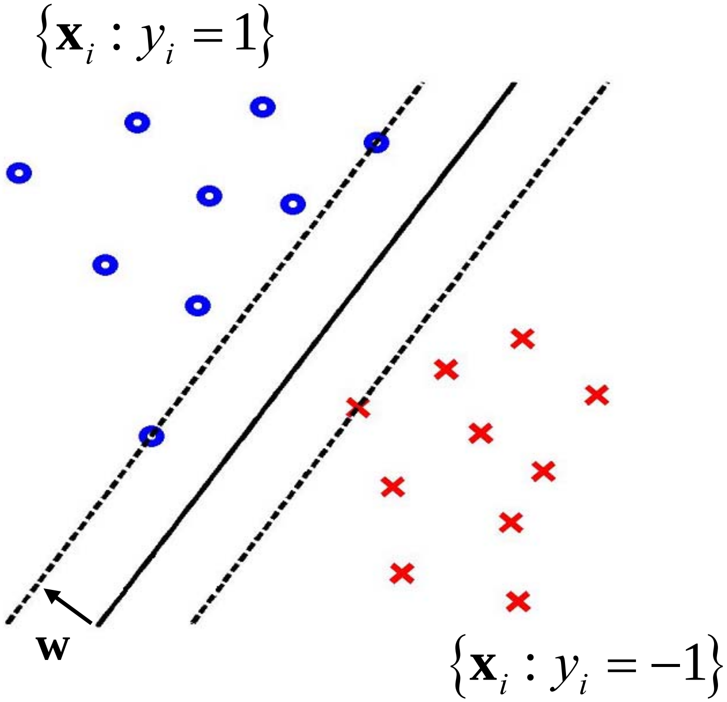




Support Vector Machines (SVMs)

Support Vector Machines are well known classifiers.

Given a dataset $\{(\mathbf{x}_i, y_i)\} \subseteq R^n \times \{\pm 1\}$



We solve the quadratic problem

$$\max_\alpha \sum_i \alpha_i - \tfrac{1}{2} \sum_{i,j} y_i y_j \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{s.t.} \quad 0 \le \alpha_i \le C, \sum_i y_i \alpha_i = 0$$

to obtain the SVM decision function

$$f(\mathbf{x}) = \sum_i \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b$$

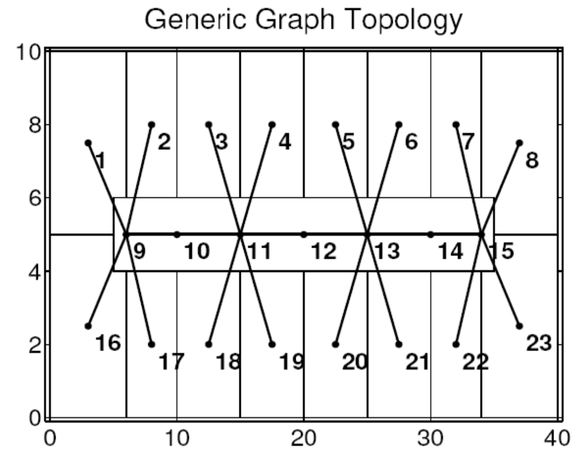
(Support Vectors are $\mathbf{x}_i$ such that $\alpha_i \neq 0$, shown as lying on dashed lines.)

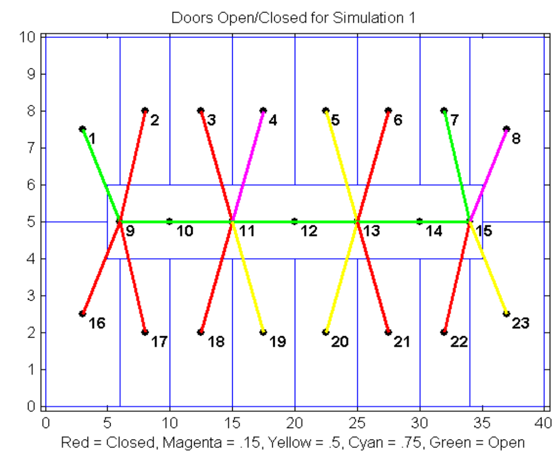
Graph Kernels

- To use SVMs with buildings, we represent building topology using graphs.
- We use weighted graphs to represent states, such as doors open/closed.
- Our SVM kernel is then a graph kernel

$$k(H_i, H_j) = \frac{1}{3} \sum_{k=1}^{3} \frac{G_i^k \cdot G_j^k}{|G_i^k||G_j^k|},$$

where $H_i = (G_1, G_2, G_3)$ is a hypergraph representing three graph states: doors, advection, and diffusion.





Building Contamination Prediction

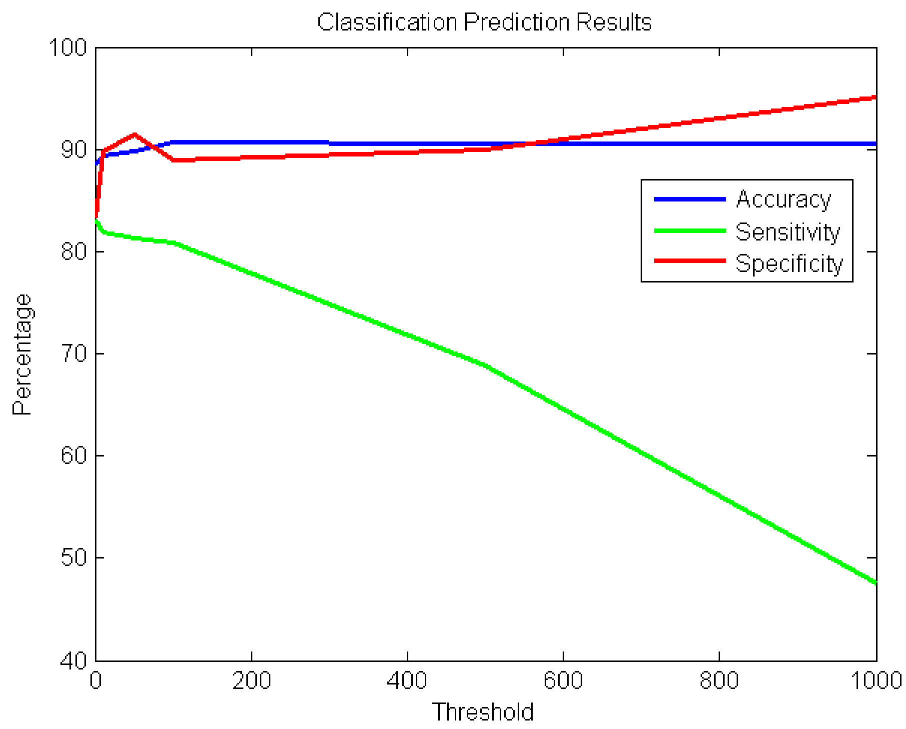
- We trained a SVM using Dataset A with 120 simulations and an invariant source location.
- We tested our predictions using 10-fold cross-validation for each room.
- For an exact contaminant prediction we used

$$q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (\hat{y}_i - \bar{y})^2}$$

  where y_i are target values, $\hat{y}_i$ are the predicted values, and $\bar{y}$ is the average target value.
- For classification prediction of contaminated vs. non-contaminated, we used accuracy, sensitivity, and specificity.

Contamination Prediction Results

- Average q^2 was 0.64 over the 23 rooms in the building.
- Accuracy was ~90% depending on threshold value for contamination.



Incorporating Partial Knowledge

- To predict source location, we need to have contaminant measurements (partial knowledge) in addition to building configuration.
- Suppose
  - σ denotes room with contaminant measurements.
  - c_i^σ denotes contaminant values in rooms σ for simulation i.
- A SVM kernel incorporating these contaminant values is given by

$$k(\mathbf{c}_i^\sigma, \mathbf{c}_j^\sigma) = \frac{\mathbf{c}_i^\sigma \cdot \mathbf{c}_j^\sigma}{|\mathbf{c}_i^\sigma||\mathbf{c}_j^\sigma|)}.$$

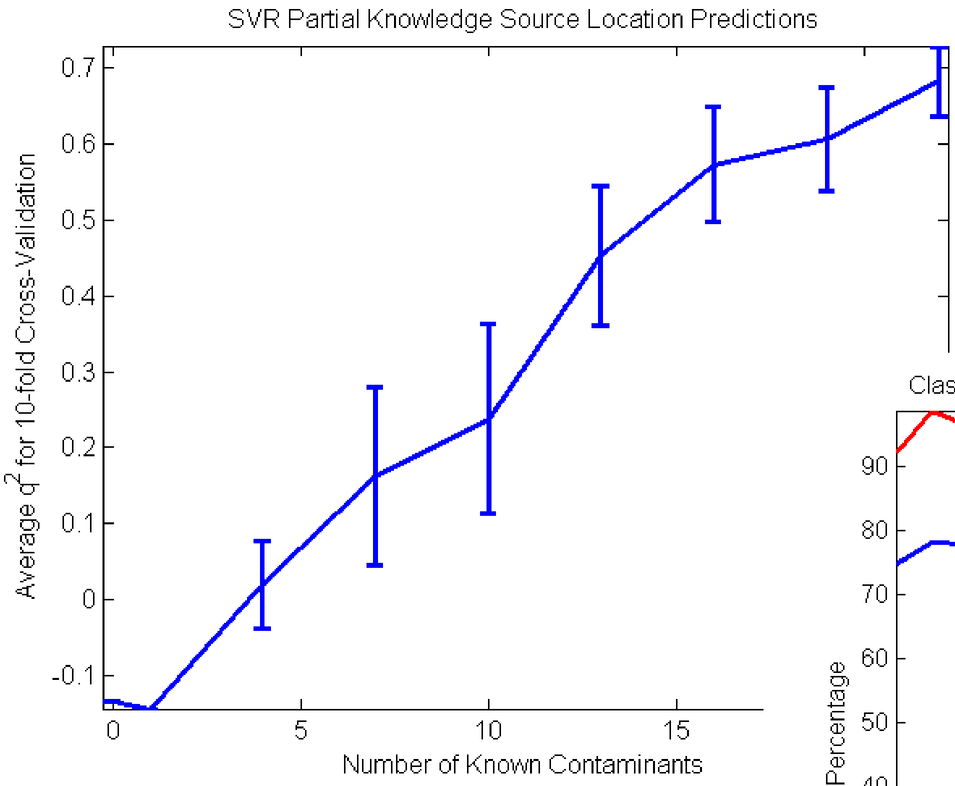
- A SVM kernel combining building configuration and contaminant values is given by

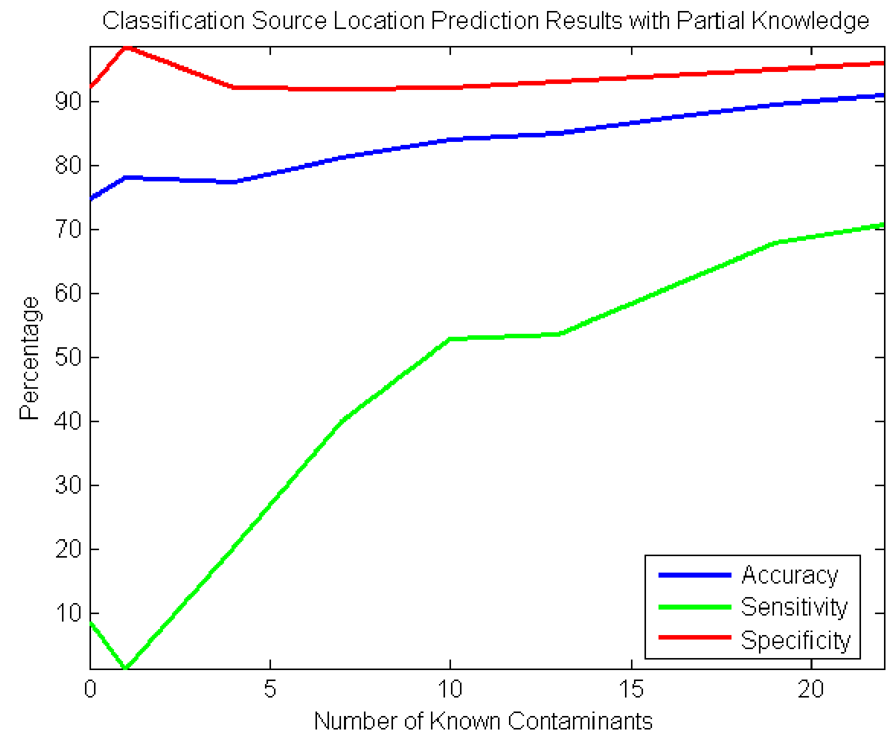
$$k((H_i, \mathbf{c}_i^\sigma), (H_j, \mathbf{c}_j^\sigma)) = \frac{1}{2}k(H_i, H_j) + \frac{1}{2}k(\mathbf{c}_i^\sigma, \mathbf{c}_j^\sigma).$$

Source Location Prediction

- We trained a SVM using Dataset B with 250 simulations and randomly varied source locations.
- We tested our predictions using 10-fold cross validation for each room.
- We used q^2 to assess our predictions of initial contaminant level in each room.
- We used accuracy, sensitivity, and specificity to assess our classification accuracy using a contaminant threshold of 0.

Source Prediction Results





Conclusions

- Demonstrated feasibility of using machine learning for modeling building contamination.
  - Requires compilation of a database of potential events for a given building.
  - Once trained, the SVM-based model is much faster than an equivalent physics-based model and is usable in real-time.
  - Can also produce SVM-based models for predicting source location.
- Future possible improvements include
  - Improve accuracy through better selection of SVM parameters.
  - Combine room predictions using structured output SVM.