Advice Refinement in Knowledge-Based Support Vector Machines

Gautam Kunapuli
University of Wisconsin–Madison
kunapg@biostat.wisc.edu

Richard Maclin
University of Minnesota–Duluth
rmaclin@d.umn.edu

Jude W. Shavlik
University of Wisconsin–Madison
shavlik@cs.wisc.edu

This research was supported by DARPA grant HR0011-07-C-0060.

Abstract. Knowledge-based support vector machines (KBSVMs) incorporate advice from domain experts to improve generalization. Imperfect advice can lead to significantly poorer models. To learn in this setting, we propose an approach that extends KBSVMs and is able to not only learn from data and advice, but also simultaneously improve the advice. The proposed approach is particularly effective for knowledge discovery in domains with few labeled examples. The additional refinement constraints are bilinear and are solved using two iterative approaches: successive linear programming and a constrained concave-convex approach.

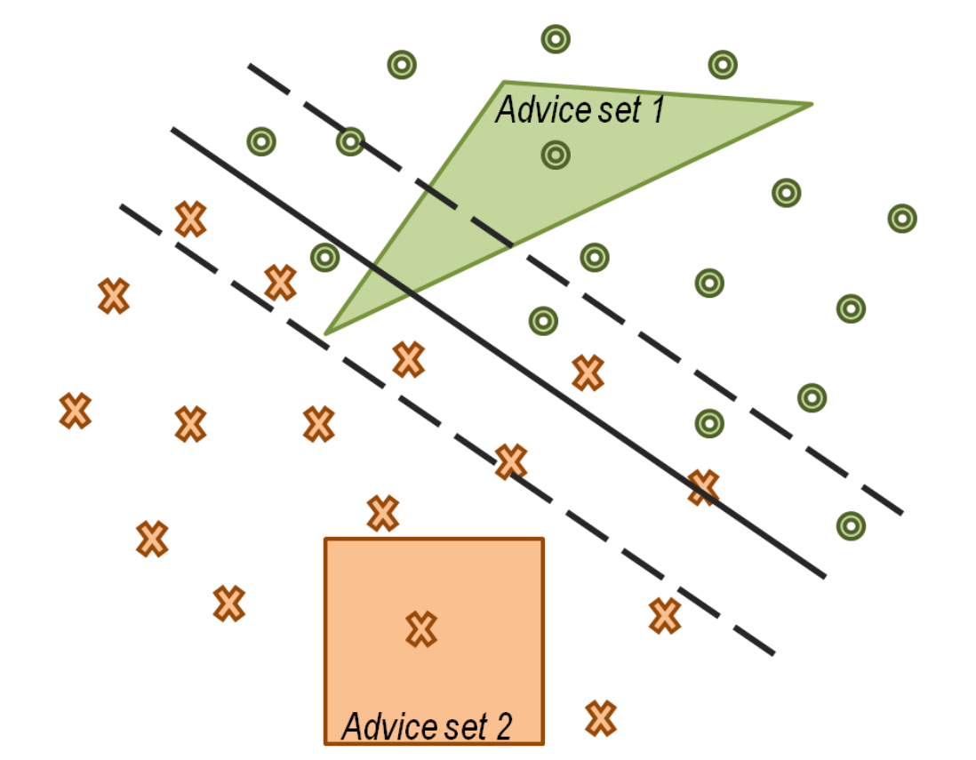
Adding Knowledge to SVMs

Polyhedral knowledge sets in the input space of data, can be added to SVMs via *knowledge-based support vector machines* (KBSVMs) [FMS03]. Knowledge sets characterize an area of input space as belonging to one of the two classes and are specified using

$$D\mathbf{x} \leq \mathbf{d} \Rightarrow z(\mathbf{w}'\mathbf{x} - \gamma) \geq 1.$$

The advice specifies that every point $\mathbf{x} \in D\mathbf{x} \leq \mathbf{d}$ lies above $\mathbf{w}'\mathbf{x} - \gamma = 1$ (if labeled $z = 1$) or below $\mathbf{w}'\mathbf{x} - \gamma = -1$ (if labeled $z = -1$).

Example: diabetes diagnosis from two features, blood glucose level and body mass index:



Expert (NIH) Advice for Type-2 Diabetes Diagnosis:

- an obsese person ($\texttt{bmi} \geq 30$) with $\texttt{gluc} \geq 126$ is at strong risk
- a person at normal weight ($\texttt{bmi} \leq 25$) with $\texttt{gluc} \leq 100$ is at low risk.

Can be expressed as polyhedral advice in input space:

$$\begin{array}{l}(\texttt{bmi} \leq 25) \wedge (\texttt{gluc} \leq 100) \Rightarrow \neg\texttt{diabetes}\\(\texttt{bmi} \geq 30) \wedge (\texttt{gluc} \geq 126) \Rightarrow \ \texttt{diabetes},\end{array}$$

In general, $D\mathbf{x} \leq \mathbf{d}$ allows linear combinations of input space features to specify advice.

Using *theorems of the alternative*, the logical implication can be reformulated as constraints with an **advice vector**, $\mathbf{u}$:

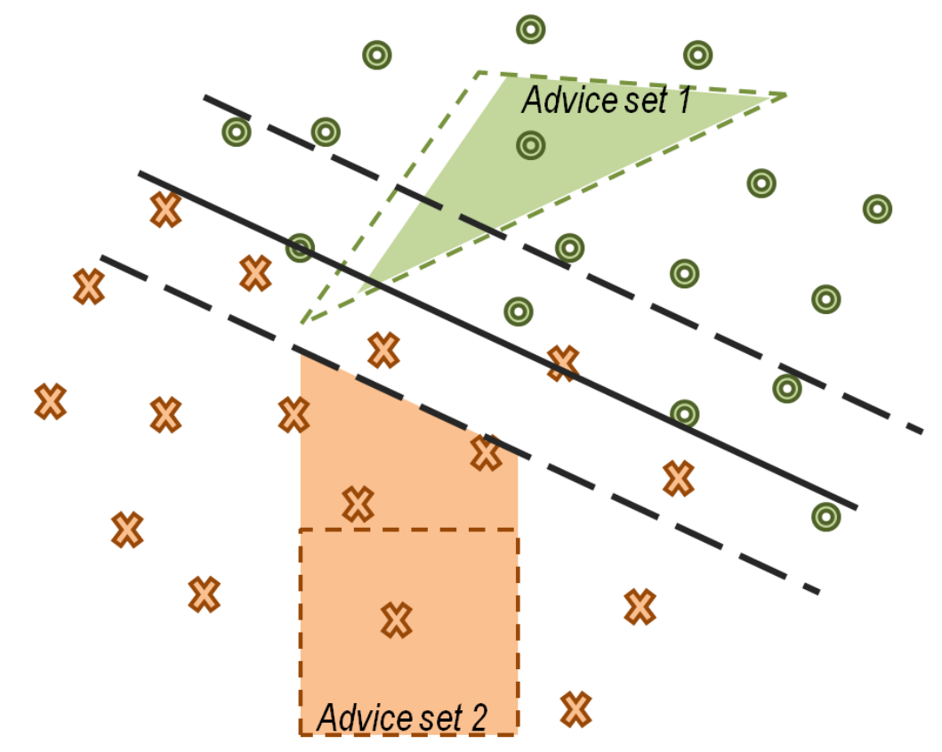
$$D'\mathbf{u} + z\,\mathbf{w} = 0, \quad -d'\mathbf{u} - z\,\gamma \geq 1, \quad \mathbf{u} \geq 0.$$

The advice vector is analogous to the dual multipliers α in SVMs: constraints of an advice set which have non-zero $\mathbf{u}^i$s are called **support constraints** and only they contribute to $\mathbf{w}$. The following is the formulation of the **knowledge-based support vector machine**:

$$\begin{array}{cl}\displaystyle\min_{\mathbf{w},b,(\boldsymbol{\xi},\mathbf{u}^i,\boldsymbol{\eta}^i,\zeta_i)\geq 0} & \|\mathbf{w}\|_1 + \lambda\mathbf{e}'\boldsymbol{\xi} + \mu\sum_{i=1}^m(\mathbf{e}'\boldsymbol{\eta}^i + \zeta_i)\\ \text{s.t.} & Y(X\mathbf{w} - b\mathbf{e}) + \boldsymbol{\xi} \geq \mathbf{e},\\ & -\boldsymbol{\eta}^i \leq D_i'\mathbf{u}^i + z_i\mathbf{w} \leq \boldsymbol{\eta}^i,\\ & -\mathbf{d}^{i\prime}\mathbf{u}^i - z_i b + \zeta_i \geq 1,\ i = 1, \ldots, m.\end{array} \tag{1}$$

Learning from Imperfect Advice

Motivation. Simultaneously learn from small data sets and refine imperfect advice provided by a domain expert.



There are two types of imperfect advice

- **Advice set 1.** Advice that lies on the wrong side of the margin is penalized as advice error through $\sum_i(\mathbf{e}'\boldsymbol{\eta}^i + \zeta_i)$ in (1).
  Refinement leads to truncation of advice.
- **Advice set 2.** Advice that lies on the correct side of the margin but far away provides no support vectors, i.e., has all its advice vectors $\mathbf{u}^i = 0$ in (1).
  Refinement leads to extension of advice.

Formulation Introduce terms $(F_i, \mathbf{f}^i)$ into the formulation that directly modify boundaries of advice regions,

$$(D_i - F_i)\mathbf{x} \leq (\mathbf{d}^i - \mathbf{f}^i) \ \Rightarrow\ z_i(\mathbf{w}'\mathbf{x} - b) \geq 1, \quad z = \pm 1$$

Extends **rule-refining support vector machines** (RRSVMs) [MWS+07] which performed refinements of the form

$$D_i\mathbf{x} \leq (\mathbf{d}^i - \mathbf{f}^i) \ \Rightarrow\ z_i(\mathbf{w}'\mathbf{x} - b) \geq 1, \quad z = \pm 1$$

Advice-Refining Knowledge-Based SVMs (arkSVMs)

Advice refinement in KBSVMs can be formulated by optimizing *model complexity + λ training error + μ advice error + ν refinement error*:

$$\begin{array}{cl}\displaystyle\min_{\mathbf{w},b,F_i,\mathbf{f}^i,(\boldsymbol{\xi},\mathbf{u}^i,\boldsymbol{\eta}^i,\zeta_i)\geq 0} & \|\mathbf{w}\|_1 + \lambda\mathbf{e}'\boldsymbol{\xi} + \mu\sum_{i=1}^m(\mathbf{e}'\boldsymbol{\eta}^i + \zeta_i) + \nu\sum_{i=1}^m\left(\|F_i\|_1 + \|\mathbf{f}^i\|_1\right)\\ \text{s.t.} & Y(X\mathbf{w} - b\mathbf{e}) + \boldsymbol{\xi} \geq \mathbf{e},\\ & -\boldsymbol{\eta}^i \leq (D_i - F_i)'\mathbf{u}^i + z_i\mathbf{w} \leq \boldsymbol{\eta}^i,\\ & -(\mathbf{d}^i - \mathbf{f}^i)'\mathbf{u}^i - z_i b + \zeta_i \geq 1, \quad i = 1, \ldots, m.\end{array} \tag{2}$$

This is the Advice-Refining KBSVM (arkSVMs).

- objective function of (2) trades-off the effect of refinement in each of the advice sets via the *refinement parameter* ν
- refining $\mathbf{d}$ alone allows only for the *translation* of the boundaries of the polyhedral advice; in arkSVMs the boundaries of the advice can be translated *and rotated*
- advice constraints are bilinear. Solved using successive linear programming (also used in [MWS+07]), and a concave-convex procedure

arkSVMs via Successive Linear Programming

Solve a sequence of linear programs while alternately fixing the bilinear variables $(F_i, \mathbf{f}^i)_{i=1}^m$ and $\{\mathbf{u}^i\}_{i=1}^m$. At the t-th iteration, the algorithm alternates between the following steps:

- **(Estimation Step)** When the refinement terms, $(\hat{F}_i^t,\ \hat{\mathbf{f}}^{i,t})_{i=1}^m$, are fixed the resulting LP becomes a standard KBSVM which attempts to find a data-estimate of the advice vectors $\{\mathbf{u}^i\}_{i=1}^m$ using the current refinement of the advice region: $(D_j - \hat{F}_j^t)\,\mathbf{x} \leq (\mathbf{d}^j - \hat{\mathbf{f}}^{j,t})$.
- **(Refinement Step)** When the advice-estimate terms $\{\hat{\mathbf{u}}^{i,t}\}_{i=1}^m$ are fixed, the resulting LP solves for $(F_i, \mathbf{f}^i)_{i=1}^m$ and attempts to further refine the advice regions based on estimates from data computed in the previous step.

Algorithm converges to a local solution.

arkSVMs via Successive Quadratic Programming

A general bilinear term $r's$, which is non-convex, can be written as the difference of two convex terms: $\frac{1}{4}\|r + s\|^2 - \frac{1}{4}\|r - s\|^2$.

The j-th component of **the bilinear advice constraint (with terms $F_i'\mathbf{u}^i$) in (2) is rewritten**:

$$D_{ij}'\mathbf{u}^i + z_i w_j - \eta_j^i + \frac{1}{4}\|F_{ij} - \mathbf{u}^i\|^2 \leq \frac{1}{4}\|F_{ij} + \mathbf{u}^i\|^2,$$

and both sides of the constraint above are convex and quadratic. We **linearize the right-hand side around an estimate of the bilinear variables $(\hat{F}_{ij}^t,\ \hat{\mathbf{u}}^{i,t})$**:

$$\begin{array}{l}D_{ij}'\mathbf{u}^i + z_i w_j - \eta_j^i + \frac{1}{4}\|F_{ij} - \mathbf{u}^i\|^2 \leq \frac{1}{4}\|\hat{F}_{ij}^t + \hat{\mathbf{u}}^{i,t}\|^2\\ \qquad\qquad + \frac{1}{2}(\hat{F}_{ij}^t + \hat{\mathbf{u}}^{i,t})'\left((F_{ij} - \hat{F}_{ij}^t) + (\mathbf{u}^i - \hat{\mathbf{u}}^{i,t})\right).\end{array}$$

Similarly, the constraint $-(D_i - F_i)'\mathbf{u}^i - z_i\mathbf{w} - \boldsymbol{\eta}^i \leq 0$, can be replaced by

$$\begin{array}{l}-D_{ij}'\mathbf{u}^i - z_i w_j - \eta_j^i + \frac{1}{4}\|F_{ij} + \mathbf{u}^i\|^2 \leq \frac{1}{4}\|\hat{F}_{ij}^t - \hat{\mathbf{u}}^{i,t}\|^2\\ \qquad\qquad + \frac{1}{2}(\hat{F}_{ij}^t - \hat{\mathbf{u}}^{i,t})'\left((F_{ij} - \hat{F}_{ij}^t) - (\mathbf{u}^i - \hat{\mathbf{u}}^{i,t})\right),\end{array}$$

while $\mathbf{d}^{i\prime}\mathbf{u}^i + z_i b + 1 - \zeta_i - \mathbf{f}^{i\prime}\mathbf{u}^i \leq 0$ is replaced by

$$\begin{array}{l}\mathbf{d}^{i\prime}\mathbf{u}^i + z_i b + 1 - \zeta_i + \frac{1}{4}\|\mathbf{f}^i - \mathbf{u}^i\|^2 \leq \frac{1}{4}\|\hat{\mathbf{f}}^{i,t} + \hat{\mathbf{u}}^{i,t}\|^2\\ \qquad\qquad + \frac{1}{2}(\hat{\mathbf{f}}^{i,t} + \hat{\mathbf{u}}^{i,t})'\left((\mathbf{f}^{i,t} - \hat{\mathbf{f}}^{i,t}) + (\mathbf{u}^i - \hat{\mathbf{u}}^{i,t})\right).\end{array}$$

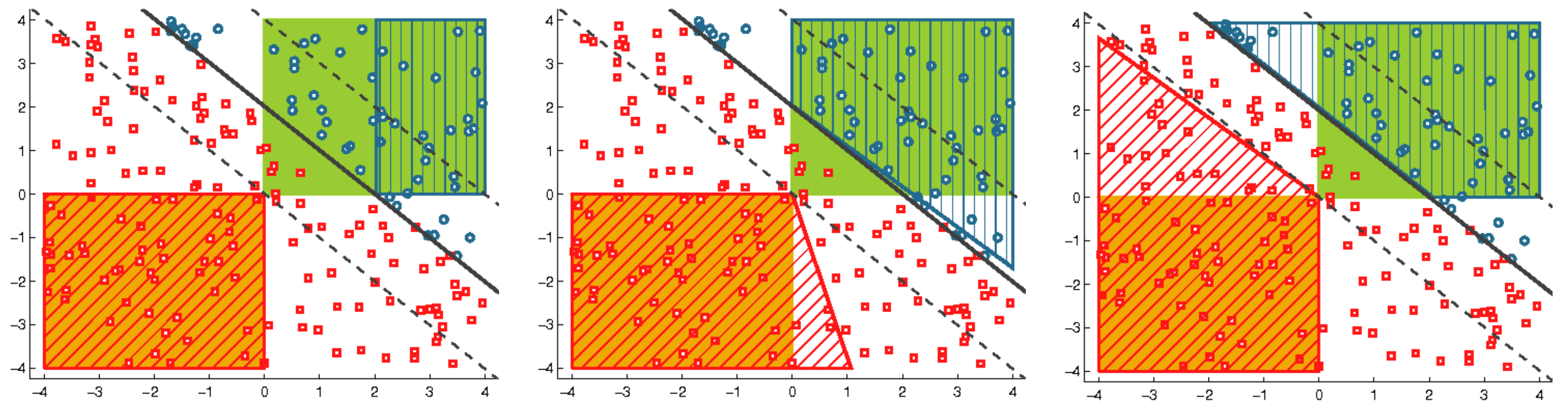
- replace original bilinear non-convex constraints of (2) with convexified relaxations gives a quadratically-constrained linear program (QCLP)
- right-hand sides are affine; entire set of constraints convex.
- feasible set of this problem is a subset of the original

Iteratively solve the QCLP: At the t-th iteration, the restricted problem uses the current estimate to construct a new feasible point and iterating this procedure produces a sequence of feasible points with decreasing objective values.

Algorithm converges to a local solution.

Toy Example

Simple 2-dimensional example consists of 200 points separated by $x_1 + x_2 = 2$. Two advice sets: $\{S_1 : (x_1, x_2) \geq 0 \ \Rightarrow\ z = +1\}$, $\{S_2 : (x_1, x_2) \leq 0 \ \Rightarrow\ z = -1\}$.



(**left**) RRSVM (**center**) arkSVM-slp (**right**) arkSVM-sqp. Orange and green unhatched regions show the original advice. The dashed lines show the margin, $\|\mathbf{w}\|_\infty$. For each method, we show the refined advice: vertically hatched for Class $+1$, and diagonally hatched for Class -1.

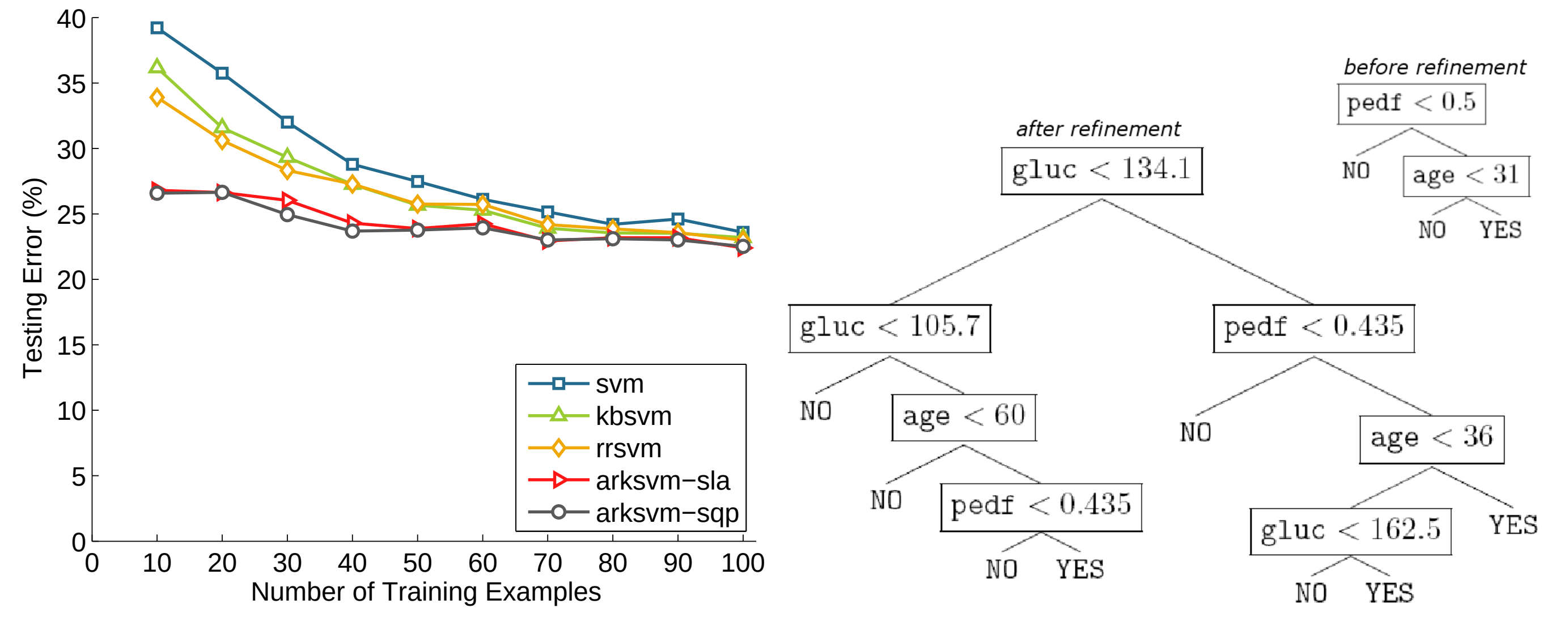
PIMA Indians Diabetes Diagnosis

Predict onset of diabetes in 768 Pima Indian women *within the next 5 years* based on eight features. Studies [HFC+98, PHK+07] show diabetes incidence among the Pima Indians is significantly higher among subjects with $\texttt{bmi} \geq 30$;. diabetes incidence is higher for a person with impaired glucose tolerance.

$$\begin{array}{lll}(\texttt{Diabetes Rule 1}) & (\texttt{gluc} \leq 126) & \Rightarrow\neg\texttt{diabetes},\\(\texttt{Diabetes Rule 2}) & (\texttt{gluc} \geq 126) \wedge (\texttt{gluc} \leq 140) \wedge (\texttt{bmi} \leq 30) & \Rightarrow\neg\texttt{diabetes},\\(\texttt{Diabetes Rule 3}) & (\texttt{gluc} \geq 126) \wedge (\texttt{gluc} \leq 140) \wedge (\texttt{bmi} \geq 30) & \Rightarrow\ \texttt{diabetes},\\(\texttt{Diabetes Rule 4}) & (\texttt{gluc} \geq 140) & \Rightarrow\ \texttt{diabetes}.\end{array}$$

The pedigree function provides a measure heredity on the subject's diabetes risk. A subject with high heredity who is at least 31 is at a significantly increased risk for diabetes in the next five years [SED+88]:

$$\begin{array}{ll}(\texttt{Diabetes Rule 5}) & (\texttt{pedf} \leq 0.5) \wedge (\texttt{age} \leq 31) \Rightarrow\neg\texttt{diabetes},\\(\texttt{Diabetes Rule 6}) & (\texttt{pedf} \geq 0.5) \wedge (\texttt{age} \geq 31) \Rightarrow\ \texttt{diabetes}.\end{array}$$



(**left**) Results averaged over 10 runs on a hold-out 412 point test set; (**right**) Approximate decision-tree showing `Diabetes Rule 6` before and after refinement (if **true** then left branch). Leaf nodes classify data according to `?diabetes`.

arkSVM algorithms produce local solutions from small data sets and imperfect expert advice. Learned models can generalize well and provide refined advice easily interpreted by domain experts.

References

[FMS03] G. Fung, O. L. Mangasarian, and J. W. Shavlik. Knowledge-based support vector classifiers. In *Advances in Neural Information Processing Systems*, volume 15, pages 521–528, 2003.

[HFC+98] M. I. Harris, K. M. Flegal, C. C. Cowie, M. S. Eberhardt, D. E. Goldstein, R. R. Little, H. M. Wiedmeyer, and D. D. Byrd-Holt. Prevalence of diabetes, impaired fasting glucose, and impaired glucose tolerance in U.S. adults. *Diabetes Care*, 21(4):518–524, 1998.

[MWS+07] R. Maclin, E. W. Wild, J. W. Shavlik, L. Torrey, and T. Walker. Refining rules incorporated into knowledge-based support vector learners via successive linear programming. In *Proc. 22nd AAAI*, pages 584–589, 2007.

[PHK+07] M. E. Pavkov, R. L. Hanson, W. C. Knowler, P. H. Bennett, J. Krakoff, and R. G. Nelson. Changing patterns of Type 2 diabetes incidence among Pima Indians. *Diabetes Care*, 30(7):1758–1763, 2007.

[SED+88] J. W. Smith, J. E. Everhart, W. C. Dickson, W. C. Knowler, and R. S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proc. Symp. Comp. Apps. and Medical Care*, pages 261–265, 1988.