Fifty-fourth Annual Allerton Conference
Allerton House, UIUC, Illinois, USA
September 27 - 30, 2016

Interpretable Machine Learning via Convex Cardinal Shape Composition

Kush R. Varshney
Mathematical Sciences Department
IBM Thomas J. Watson Research Center
Yorktown Heights, New York 10598
Email: krvarshn@us.ibm.com

***Abstract*—For safety reasons, the interpretability of models is important in consequential applications of supervised classification in which predictions are used to support human decision makers. In this paper, we extend cardinal shape composition, a new method developed in the image processing and computer vision literature for image segmentation, to general machine learning problems. Our transformation results in a computationally-tractable ℓ_1-regularized hinge loss optimization over a shape dictionary. This approach yields human-interpretable models with an appropriate choice of atomic shapes in the dictionary used to compose decision boundaries.**

I. Introduction

Predictions from models induced by machine learning algorithms are increasingly being used to support decisions by people in domains of consequence, e.g. medical treatment, prison parole, loan approval, and job promotion [1]. In such applications, it is critical to control the probability of expected harms and the possibility of unexpected harms. This amounts to introducing safety into the decision-making system by minimizing both risk and epistemic uncertainty. One key approach for minimizing epistemic uncertainty in machine learning, utilizing the safety strategy known as inherently safe design [2], is insisting on models that are comprehensible, transparent, explainable, and interpretable to humans [3].

In contrast to the learning of black-box models such as large ensembles, deep neural networks, and complicated kernel machines, interpretable machine learning is focused on formats such as rule sets, scorecards and decision trees that can be comprehended by people [4], [5]. The goal of risk minimization still holds in learning the models, and thus training should be done with as accurate generalization as possible; it is through the constrained model format that epistemic uncertainty minimization is achieved.

Interpretable machine learning has many facets [6]: one can consider global interpretability in which the entire model is transparent, instance-level interpretability in which explanations can be derived for individual test samples, or visualizations that let the user understand a model by interacting with it. One can consider directly learning interpretable models from training data or post hoc interpretations of learned complicated models. Directly-learned global interpretability is most consistent with inducing inherently safe design and is what we focus on in this work. Our focus is also on rules and rule sets or lists for the supervised classification problem.

Many of the older methods in the artificial intelligence literature are interpretable [7]–[9]. The older interpretable model learning algorithms are generally greedy or heuristic in nature and usually have inferior predictive performance to newer uninterpretable approaches. However, recent work is revisiting the problem of interpretable learning and achieving predictive performance approaching that of uninterpretable methods. Some approaches involve Bayesian statistics [10]–[14], others involve mixed integer programming [15]–[18], and others build upon sparse signal representation and compressed sensing [19]–[23].

In this paper, we propose a method for interpretable supervised machine learning that takes a new method for image segmentation and other related image processing and computer vision tasks, and reimagines it as a way to learn a classifier for general (non-image) data. This new image segmentation method, convex cardinal shape composition [24], [25], is a convex optimization and sparse signal representation perspective on the rich tradition of active contours and level set methods for image segmentation [26]–[28]. Starting from a large dictionary of possible atomic shapes (constructed using prior knowledge), the method composes a larger shape from a sparse subset of dictionary elements through union, intersection, and set difference operations implemented using linear combinations of characteristic functions of the shapes and the Heaviside function. The naturally nonconvex segmentation problem in the formulation is convexified using hinge loss-like expressions and regularized by an ℓ_1 constraint on the coefficient vector of the dictionary elements.

Our transfer of the image processing problem to the general supervised machine learning problem is accomplished by altering the main energy functional of the objective from indicating segmentation quality to being the average loss on the provided training data set. In particular, taking advantage of the hinge loss convexification already present in the convex cardinal shape composition method, the training loss is the hinge loss. We can achieve interpretable models — our main interest in this work — with an appropriate choice of shape dictionary. We can learn Boolean rules, a set of small hypercubes (similar in form to [17]), and other interpretable model forms. With other choices of shape dictionaries, we can recover the uninterpretable ℓ_1-regularized kernel support vector machine (SVM) [29]. Thus our proposed method is a generalization of

978-1-5090-4550-1/16/$31.00 ©2016 IEEE

the SVM that can yield both interpretable and uninterpretable models and anything on the continuum in-between, depending on the choice of shape dictionary. In all cases, the learning is a convex optimization problem.

We have successfully followed the general practice of converting an image segmentation problem into a supervised classification problem before [30], [31].

II. Convex Cardinal Shape Composition

In the image segmentation problem solved by convex cardinal shape composition, we have an image $I(x)$ defined on a domain $\Omega \subset \mathbb{R}^d$ with pixel locations $x \in \Omega$. Our goal is to partition Ω into the foreground region Σ (not necessarily simply connected) and background region $\Omega \setminus \Sigma$ in a way that minimizes some functional based on $I(x)$ and some regularization term.

Let $\Pi_{\text{fg}}(x)$ and $\Pi_{\text{bg}}(x)$ be known and fixed functions that characterize how well $I(x)$ matches the appearance of the foreground and background, respectively. Then an unregularized image segmentation problem is:

$$\min_{\Sigma} \int_{\Sigma} \Pi_{\text{fg}}(x)dx + \int_{\Omega\setminus\Sigma} \Pi_{\text{bg}}(x)dx. \tag{1}$$

Now specific to the convex cardinal shape composition formulation, we have a dictionary of shapes $\mathfrak{D} = \{\mathcal{S}_1, \dots, \mathcal{S}_m\}$, where each $\mathcal{S}_i \subset \Omega$. A shape $\mathcal{S}_i$ has characteristic function $\chi_{\mathcal{S}_i}(x)$ defined as:

$$\chi_{\mathcal{S}_i}(x) = \begin{cases} 1, x \in \mathcal{S}_i \\ 0, x \notin \mathcal{S}_i. \end{cases} \tag{2}$$

A larger shape can be composed by a linear combination of shapes from the dictionary as:

$$L_{\boldsymbol{\alpha}}(x) = \sum_{i=1}^{m} \alpha_i \chi_{\mathcal{S}_i}(x), \tag{3}$$

where α_i are coefficients. Letting $H(\cdot)$ be the Heaviside unit step function that takes value one for non-negative inputs and value zero for negative inputs, $H(L_{\boldsymbol{\alpha}}(x))$ is the characteristic function of the composed shape. Through this representation, the composed shape can involve any union and set difference operation among the dictionary atoms. Moreover, the segmentation problem formulation may be written as:

$$\min_{\boldsymbol{\alpha}} \int_{\Omega} \left(\Pi_{\text{fg}}(x) - \Pi_{\text{bg}}(x)\right) H(L_{\boldsymbol{\alpha}}(x))dx. \tag{4}$$

The problem (4), however, is not convex. As shown in [24], it may be convexified as:

$$\begin{aligned}\min_{\boldsymbol{\alpha}} \int_{\Omega} &\left(\Pi_{\text{fg}}(x) - \Pi_{\text{bg}}(x)\right)_+ \max(0, L_{\boldsymbol{\alpha}}(x))dx \\ &- \int_{\Omega} \left(\Pi_{\text{bg}}(x) - \Pi_{\text{fg}}(x)\right)_+ \min(1, L_{\boldsymbol{\alpha}}(x))dx,\end{aligned} \tag{5}$$

where the $(\cdot)_+$ operator returns zero if its argument is negative and the argument itself if it is non-negative. The expressions inside the integrals have the form of hinge functions.

The final convex cardinal shape composition formulation further imposes an ℓ_1 constraint on $\boldsymbol{\alpha}$ for sparsity:

$$\begin{aligned}\min_{\|\boldsymbol{\alpha}\|_1 \leq \tau} \int_{\Omega} &\left(\Pi_{\text{fg}}(x) - \Pi_{\text{bg}}(x)\right)_+ \max(0, L_{\boldsymbol{\alpha}}(x))dx \\ &- \int_{\Omega} \left(\Pi_{\text{bg}}(x) - \Pi_{\text{fg}}(x)\right)_+ \min(1, L_{\boldsymbol{\alpha}}(x))dx,\end{aligned} \tag{6}$$

The problem can be efficiently optimized using either linear programming or the alternating direction method of multipliers [25].

III. Supervised Classification Problem

In this section, we adapt (6) to the supervised binary classification problem. In this machine learning problem, we have features $x \in \Omega \subset \mathbb{R}^d$ and labels $y \in \{-1, +1\}$. We are given training samples $\{(x_1, y_1), \dots, (x_n, y_n)\}$ from which to learn a classifier $\hat{y} : \Omega \to \{-1, +1\}$. A classifier is also a partitioning of the domain Ω into Σ and $\Omega \setminus \Sigma$.

The classifier can also be defined based on a decision function $\varphi(x)$ that takes negative values in Σ and positive values in $\Omega \setminus \Sigma$. Then $\hat{y}(x) = 2H(\varphi(x)) - 1$. The zero level set of $\varphi(x)$ is the decision boundary separating the two classes. Margin, or distance away from the decision boundary signed such that incorrect classifications take negative values and correct classifications take positive values, can be represented as $y_j\varphi(x_j)$ due to the fact that the class labels are in the set $\{-1, +1\}$ and multiplication of these values is positive when the same sign and negative when different signs.

A common approach for learning $\hat{y}$ (in primal form) is to minimize an empirical margin-based loss of the training data:

$$\frac{1}{n}\sum_{j=1}^{n} \ell\left(y_j \varphi_j(x_j)\right), \tag{7}$$

with an additional constraint for regularization. Common margin-based loss functions $\ell(\cdot)$ include logistic loss, exponential loss, and hinge loss:

$$\begin{aligned}&\ell_{\text{hinge}}\left(y\varphi(x)\right) \triangleq \\ &\max\left(0, 1 - y\varphi(x)\right) = \left(1 - y\varphi(x)\right)_+ = 1 - \min\left(1, y\varphi(x)\right).\end{aligned} \tag{8}$$

In contrast to the image segmentation problem in which we are dealing with the continuous field $I(x)$, we only have the point samples $\{x_1, \dots, x_n\}$ in the supervised learning problem. Also, due to the structure of the margin-based loss function, we do not require separate functions to indicate goodness of fit in the Σ and $\Omega \setminus \Sigma$ regions: one is enough. Utilizing the Dirac delta function and making appropriate choices including letting $\varphi(x) = L_{\boldsymbol{\alpha}}(x)$, we may specialize (6) to obtain:

$$\begin{aligned}&\min_{\|\boldsymbol{\alpha}\|_1 \leq \tau} \int_{\Omega} \sum_{j=1}^{n} \left(1 - yL_{\boldsymbol{\alpha}}(x)\right)_+ \delta(x - x_j)dx \\ &\qquad = \min_{\|\boldsymbol{\alpha}\|_1 \leq \tau} \frac{1}{n}\sum_{j=1}^{n} \left(1 - y_jL_{\boldsymbol{\alpha}}(x_j)\right)_+ .\end{aligned} \tag{9}$$

This formulation has the same convexification using hinge functions; we can solve this problem using the convex cardinal shape composition optimization machinery.

IV. Choice of Shape Dictionary

The development in Section III formulates the supervised classification machine learning problem, but glosses over the choice of shape dictionary $\mathfrak{D}$. Recall that $L_{\boldsymbol{\alpha}}(x) = \sum_{i=1}^{m} \alpha_i \chi_{\mathcal{S}_i}(x)$. Our goal in this work is interpretability, and thus the shapes, which are atoms to compose the decision boundary, should be easily comprehensible.

Interpretable models, including rules, tend to have simple axis-aligned splits as decision boundaries composed with simple Boolean expressions. Boolean OR-rules can be recovered in our proposed model by choosing the shapes $\mathcal{S}_i$ as axis-aligned half-spaces and imposing an extra non-negativity constraint on the elements of $\boldsymbol{\alpha}$. Without that non-negativity constraint, more interesting but still also interpretable decision rules can be obtained.

Another type of atomic shape that may be considered is small hypercubes. Especially relevant for imbalanced data and learning so-called box drawing models [17], a sparse composition of hypercubes allows for interpretability by highlighting pockets of a particular class label in an easy-to-understand axis-aligned way. Axis-aligned slabs that extend infinitely in some dimensions are also a possible interpretable shape dictionary.

Other choices for $\mathfrak{D}$ will make the resulting classifiers uninterpretable. As such an example, consider complicated kernel machines cited as uninterpretable in Section I. Our proposed formulation is related to kernel machines as follows. The primal form of the ℓ_1-regularized kernel SVM is the same as (9) if we replace $L_{\boldsymbol{\alpha}}(x) = \sum_{i=1}^{m} \alpha_i \chi_{\mathcal{S}_i}(x)$ with $\sum_{j=1}^{n} \alpha_j y_j K(x, x_j)$, where K is a kernel function [29]. Compactly-supported kernels could be considered as shapes in our proposed formulation [32], but would not, in general, lead to interpretable compositions.

As such, the proposed methodology for machine learning based on convex cardinal shape composition presents a way to tractably learn models anywhere on the continuum between interpretability and uninterpretability depending on the choice of shape dictionary. Simple axis-aligned shapes lead to interpretable models and more complicated shapes lead to uninterpretable ones.

V. Conclusion

In this early research, we have proposed a new methodology for interpretable machine learning built upon the foundation of the image segmentation algorithm known as convex cardinal shape composition. It follows a different paradigm than existing new developments in interpretable learning such as ones built upon Bayesian inference, mixed integer programming, and Boolean compressed sensing. The interpretability of the model enters through the choice of atoms by which the decision boundary of the overall model is constructed: axis-aligned hypercubes, slabs, and half-spaces tend to lead to interpretable models. The formulation we develop has a close relationship to the ℓ_1-regularized SVM, but differs due to the shape dictionary. There is much opportunity to further develop this initial proposal in many directions, both theoretically and empirically.

References

[1] K. R. Varshney, "Data science of the people, for the people, by the people: A viewpoint on an emerging dichotomy," in *Proc. Data for Good Exchange Conf.*, New York, NY, Sep. 2015.

[2] N. Möller and S. O. Hansson, "Principles of engineering safety: Risk and uncertainty reduction," *Reliab. Eng. Syst. Safe.*, vol. 93, no. 6, pp. 798–805, Jun. 2008.

[3] K. R. Varshney, "Engineering safety in machine learning," in *Proc. Inf. Theory Appl. Workshop*, La Jolla, CA, Feb. 2016.

[4] A. A. Freitas, "Comprehensible classification models – a position paper," *SIGKDD Explorations*, vol. 15, no. 1, pp. 1–10, Jun. 2013.

[5] C. Rudin, "Algorithms for interpretable machine learning," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, New York, NY, Aug. 2014, p. 1519.

[6] B. Kim, D. M. Malioutov, and K. R. Varshney, Eds., *Proc. ICML Workshop Human Interpretability Mach. Learn. (WHI)*, New York, NY, Jun. 2016.

[7] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man-Mach. Studies*, vol. 27, no. 3, pp. 221–234, Sep. 1987.

[8] R. L. Rivest, "Learning decision lists," *Mach. Learn.*, vol. 2, no. 3, pp. 229–246, Nov. 1987.

[9] W. W. Cohen, "Fast effective rule induction," in *Proc. Int. Conf. Mach. Learn.*, Tahoe City, CA, Jul. 1995, pp. 115–123.

[10] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *Ann. Appl. Stat.*, vol. 9, no. 3, pp. 1350–1371, Sep. 2015.

[11] F. Wang and C. Rudin, "Falling rule lists," in *Proc. Int. Conf. Artif. Intell. Stat.*, San Diego, CA, May 2015, pp. 1013–1022.

[12] T. Wang, C. Rudin, F. Doshi-Velez, Y. Liu, E. Klampfl, and P. MacNeille, "Or's of and's for interpretable classification, with application to context-aware recommender systems," http://arxiv.org/pdf/1504.07614.pdf, Apr. 2015.

[13] Ş. Ertekin and C. Rudin, "A Bayesian approach to learning scoring systems," *Big Data*, vol. 3, no. 4, pp. 267–276, Dec. 2015.

[14] H. Yang, C. Rudin, and M. Seltzer, "Scalable Bayesian rule lists," http://arxiv.org/pdf/1602.08610.pdf, Feb. 2016.

[15] E. Boros, P. L. Hammer, T. Ibaraki, and A. Kogan, "Logical analysis of numerical data," *Math. Program.*, vol. 79, no. 1, pp. 163–190, Oct. 1997.

[16] D. Bertsimas, A. Chang, and C. Rudin, "An integer optimization approach to associative classification," in *Adv. Neur. Inf. Process. Syst. 25*, 2012, pp. 269–277.

[17] S. T. Goh and C. Rudin, "Box drawings for learning with imbalanced data," in *Proc. ACM SIGKDD Conf. Knowl. Discov. Data Min.*, New York, NY, Aug. 2014, pp. 333–342.

[18] B. Ustun and C. Rudin, "Supersparse linear integer models for optimized medical scoring systems," *Mach. Learn.*, vol. 102, no. 3, pp. 349–391, Mar. 2016.

[19] M. Marchand and J. Shawe-Taylor, "The set covering machine," *J. Mach. Learn. Res.*, vol. 3, pp. 723–746, Dec. 2012.

[20] U. Rückert and S. Kramer, "Margin-based first-order rule learning," *Mach. Learn.*, vol. 70, no. 2–3, pp. 189–206, Mar. 2008.

[21] D. M. Malioutov and K. R. Varshney, "Exact rule learning via Boolean compressed sensing," in *Proc. Int. Conf. Mach. Learn.*, Atlanta, GA, Jun. 2013, pp. 765–773.

[22] A. Emad, K. R. Varshney, and D. M. Malioutov, "A semiquantitative group testing approach for learning interpretable clinical prediction rules." in *Proc. Signal Process. Adapt. Sparse Struct. Repr. Workshop*, Cambridge, UK, Jul. 2015.

[23] G. Su, D. Wei, K. R. Varshney, and D. M. Malioutov, "Learning sparse two-level Boolean rules," in *Proc. IEEE Workshop Mach. Learn. Signal Process.*, Vietri Sul Mare, Italy, Sep. 2016.

[24] A. Aghasi and J. Romberg, "Convex cardinal shape composition," *SIAM J. Imaging Sci.*, vol. 8, no. 4, pp. 2887–2950, 2015.

[25] ——, “Learning shapes by convex composition,” http://arxiv.org/pdf/1602.07613.pdf, Jul. 2016.

[26] T. F. Chan and L. A. Vese, “Active contours without edges,” *IEEE Trans. Image Process.*, vol. 10, no. 2, pp. 266–277, Feb. 2001.

[27] A. Tsai, A. Yezzi, and A. S. Willsky, “Curve evolution implementation of the Mumford-Shah functional for image segmentation, denoising, interpolation, and magnification,” *IEEE Trans. Image Process.*, vol. 10, no. 8, pp. 1169–1186, Aug. 2001.

[28] N. Paragios and R. Deriche, “Geodesic active regions and level set methods for supervised texture segmentation,” *Int. J. Comput. Vis.*, vol. 46, no. 3, pp. 223–247, Feb. 2002.

[29] J. Zhu, S. Rosset, R. Tibshirani, and T. J. Hastie, “1-norm support vector machines,” in *Adv. Neur. Inf. Process. Syst. 16*, Dec. 2003, pp. 49–56.

[30] K. R. Varshney and A. S. Willsky, “Classification using geometric level sets,” *J. Mach. Learn. Res.*, vol. 11, pp. 491–516, Feb. 2010.

[31] T. Lin, H. Xue, L. Wang, B. Huang, and H. Zha, “Supervised learning via Euler’s elastica models,” *J. Mach. Learn. Res.*, vol. 16, pp. 3637–3686, Dec. 2015.

[32] M. G. Genton, “Classes of kernels for machine learning: A statistics perspective,” *J. Mach. Learn. Res.*, vol. 2, pp. 299–312, Dec. 2001.