Computational Geometry Neighborhoods for Local Learning

Maya R. Gupta (gupta@ee.washington.edu), Eric K. Garcia, Yihua Chen
University of Washington, Seattle, WA

Local learning uses only a neighborhood of training samples to make predictions for a test sample, and includes popular learning techniques such as the k-NN classifier and local linear regression, as well as more recent algorithms, such as hyperplane k-NN [1], SVM-KNN [2], and local similarity discriminant analysis [3]. A key issue in local learning that has been relatively unstudied is the choice of neighborhood. It is standard to use the nearest k neighbors, or all the neighbors in some hypersphere around the test sample. Such definitions do not ensure that the neighbors will be well spread-out around the test point, and generally require cross-validation of a neighborhood size parameter. Ideas from computational geometry offer neighborhood definitions that adapt automatically to the spatial distribution of the feature vectors. Such neighborhood definitions can adaptively size the neighborhood and may have provably optimal properties.

Perhaps the first work in this area is Sibson's natural neighbors, proposed for linear interpolation [4]. The natural neighbors are defined by the Voronoi tessellation $\mathcal{V}$ of the training set and test point. Given $\mathcal{V}$, the natural neighbors of a test point g are defined to be those training points $\{x_j\}$ whose Voronoi cells are adjacent to the cell containing g. An example of the natural neighbors is shown in the left diagram of Fig. 1. Though proposed for linear interpolation, the natural neighbors could be used as a neighborhood definition for any local learning method. Another computational geometric approach to local learning is to use the Gabriel neighbors [5, pg.90][6].

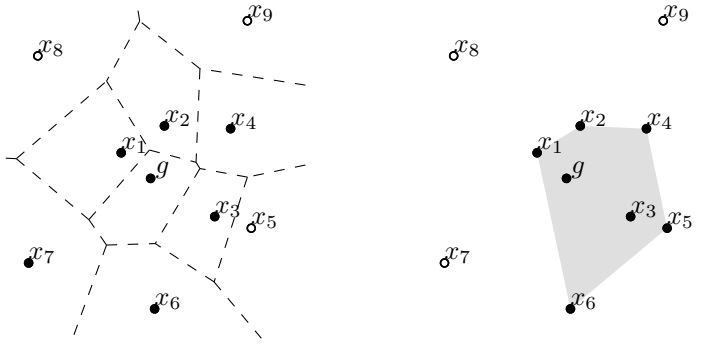
Recently, we have proposed that an effective neighborhood for a test point would include the test point in the convex hull of the neighbors [7], [8]. It is easy to show that the natural neighbors satisfy this property. We proposed another enclosing neighborhood called *enclosing k-NN*: the first k neighbors whose convex hull closure is closer to the test sample [8]; an example is shown in the right diagram of Fig. 1.

We showed that on average, asymptotically, the enclosing k-NN will consist of $2d + 1$ neighbors, where d is the dimension of the space. We proved that linear regression on an *enclosing neighborhood* has a bounded estimation variance, if the true function is a linear

Two enclosing neighborhoods for g.

Left: Natural neighbors marked with solid circles.

Right: Enclosing k-NN neighbors marked with solid circles.



$\mathcal{J}_g^{\mathrm{NN}} = \{x_1, x_2, x_3, x_4, x_6, x_7\}$ $\mathcal{J}_g^{ek\mathrm{NN}} = \{x_1, \ldots, x_6\}$

trend with additive noise [8]. Experimental results demonstrated that local linear regression over enclosing neighborhoods can significantly significantly reduce color management error over the state-of-the-art.

When the test sample cannot be enclosed in the convex hull of the set of training samples, it is less clear what the value of any particular enclosing neighborhood is. The spatial distribution of the training samples should still be useful for deciding what neighbors to learn from, but how to do this is an open question.

Some learning problems require making predictions given only pairwise similarities (or dissimilarities) between samples. Local learning methods can be applied to these problems as well [3]. Here too, adaptive neighborhood definitions are needed that will ensure neighborhoods contain samples that are highly similar to the test sample, but also a diversity of highly similar training samples. To this end, assume that $\mathcal{B}$ is some enumerated abstract space of samples, n labeled training samples $\{x_j\} \in \mathcal{B}$, test sample $g \in \mathcal{B}$ and suppose there is any similarity function $s : \mathcal{B} \times \mathcal{B}$. We define a similarity-based Voronoi tessellation such that the Voronoi cell of x_j is $V_j = \{z \in \mathcal{B} | s(z, x_j) > s(z, x_i)\}$ for all $i \neq j$. Then we define the similarity-based natural neighbors as those training samples whose Voronoi cell changes if the Voronoi tessellation is performed on the set $\{g, x_1, x_2, \ldots, x_n\}$. We will show how these definitions may be used in practice for semi-supervised similarity-based classification.

References

[1] P. Vincent and Y. Bengio, "K-local hyperplane and convex distance nearest neighbor algorithms," *NIPS*, pp. 985–992, 2001.

[2] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "SVM-KNN: discriminative nearest neighbor classification for visual category recognition," *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 2126–2136, 2006.

[3] L. Cazzanti and M. R. Gupta, "Local similarity discriminant analysis," *Proc. Intl. Conf. on Machine Learning*, 2007.

[4] R. Sibson, *Interpreting multivariate data*. John Wiley, 1981, ch. A brief description of natural neighbour interpolation, pp. 21–36.

[5] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag Inc., 1996.

[6] B. Bhattacharya, K. Mukherjee, and G. Toussaint, "Geometric decision rules for instance-based learning," *Lecture Notes Computer Science*, pp. 60–69, 2005.

[7] M. R. Gupta, "Custom color enhancements," *Proc. of the IEEE Intl. Conf. on Image Processing*, pp. 968–971, 2005.

[8] M. R. Gupta, E. K. Garcia, and E. M. Chin, "Adaptive local linear regression with application to printer color management," *IEEE Trans. on Image Processing*, 2007, to appear.