Optimally Quantized and Smoothed Histograms

Mingzhou Song and Robert M. Haralick

Department of Electrical Engineering, University of Washington, Seattle, WA 98195-2500
E-mail: msong@u.washington.edu, haralick@ee.washington.edu

Abstract

We propose an approach using optimal quantization and smoothing to generate adaptive histograms for multi-class one dimensional data. The discretization of data is optimal in maximizing a quantizer performance measure, defined by a combination of average log likelihood, entropy and correct classification probability. The optimal partition is found by dynamic programming. The density of each bin is obtained by a smoothing technique that can be considered a generalized k nearest neighbor density estimation algorithm. However, our smoothing approach is much more efficient. Experimental results demonstrated the effectiveness of the optimally quantized and smoothed histograms. Even though obtaining one takes about quadratic time in sample size, an optimal histogram is much more efficient to use than typical kernel methods. Therefore, optimal histograms are more suitable in applications with massive data set.

1. Introduction

Equal bin width histograms are widely used because they are relatively efficient to obtain and apply. However, they are not statistically efficient for two reasons. First, the bins are blindly allocated, not adapting to the data. Second, the normalized frequency may be zero for many bins and there is no guarantee for the consistency of the density estimates.

We advocate data based variable bin width histograms and all bins should contain non-zero density estimates. [8, 9] have proven the consistency of data driven histogram density estimates and proposed general partition schemes. [11] suggests a variable bin width histogram by inversely transforming an equal bin width histogram obtained on the transformed data. However, it is not clear how the transform function should be chosen in general.

A better approach is to find a partition scheme by optimizing a quantizer measure. Entropy and likelihood [4, 6] have been proposed as quantizer measures. Discretization of multi-class 1-D data is studied extensively in [2, 3, 7]. They use measures quantifying the discrimination power of quantizers. In [3], dynamic programming is used to find the optimal partition based on additive measures. [1] improves the practical efficiency of the algorithm on well-behaved additive measures, asserting that many widely used classification measures are well-behaved. Smoothing makes histograms visually appealing and suitable for application. WARPing [5] and averaging shifted histograms [10] methods smooth histograms. Otherwise there are relatively few histogram smoothing methods. Smoothing is important in maintaining consistency of histogram density estimates and deserves further study.

In Section 2, we define the quantizer performance measure. In Section 3, we introduce a dynamic programming algorithm to find an optimal partition. In Section 4, we propose an algorithm to obtain smooth histogram density estimates. In Section 5, we show some experiment results to demonstrate the effectiveness of optimal histograms.

2. Quantizer performance measure

Let $\mathcal{X}_N$ be a sample of size N and $\mathcal{Y}_N$ be the corresponding class labels. Let K be the total number of classes and $N(y)$ be the total number of class y data. Let Q be a quantizer with L bins. Let $\Delta(q)$ be the width of bin q. Let N_q be the total number of data in bin q. Let $N_q(y)$ be the total number of class y data in bin q.

Average log likelihood. Kullback-Leibler divergence from $\hat{p}(x)$ to $p(x)$ is

$$D(p||\hat{p}) = \int p(x) \log \frac{p(x)}{\hat{p}(x)} dx = \mathbf{E}[\log p(X)] - \mathbf{E}[\log \hat{p}(X)]$$

which, being non-negative (zero only when $\hat{p}(x) = p(x)$), should be minimized. As $p(x)$ is fixed, maximizing $\mathbf{E}[\log \hat{p}(X)]$ is equivalent to minimizing $D_{KL}(p||\hat{p})$. Let $p(q|y)$ be the density of bin q. Then $\mathbf{E}[\log \hat{p}(X|Y)]$ can be estimated by $\frac{1}{N(y)} \log \prod_{q=1}^{L} (p(q|y))^{N_q(y)}$. The *overall*

average log likelihood of a quantizer Q is

$$J(Q) = \frac{1}{N}\sum_{y=1}^{K} N(y)\mathbf{E}[\log \hat{p}(X|y)] = \frac{1}{N}\sum_{y=1}^{K} \log \prod_{q=1}^{L} (p(q|y))^{N_q(y)}$$

When the class number ratio $N(1) : N(2) : \cdots : N(K)$ is representative for the true data, the overall average log likelihood is preferred, with the log likelihood of popular classes being emphasized.

The *mean class average log likelihood* is

$$J(Q) = \frac{1}{K}\sum_{y=1}^{K} \mathbf{E}[\log \hat{p}(X|y)] = \frac{1}{K}\sum_{y=1}^{K} \frac{\log \prod_{q=1}^{L} (p(q|y))^{N_q(y)}}{N(y)}$$

When the class number $N(y)$ is randomly decided or every class has equal importance, the mean class average log likelihood is preferred, with every class contributing equally to the log likelihood of the quantizer.

Correct classification probability. Let $P(y)$ be the prior probability of class Y. Within bin q, the Bayes' rule is equivalent to $y_q^* = \operatorname{argmax}_y P(y)N_q(y)/N(y)$. Let $N_c(q)$ be the number of correct decisions in bin q, i.e.,$N_c(q) = N_q(y_q^*)$. We give the definition of the correct classification probability in two situations. The *overall correct classification probability* is

$$P_c(Q) = \frac{\sum_{q=1}^{L} N_c(q)}{N} \tag{1}$$

The *mean class correct classification probability* is

$$P_c(Q) = \frac{1}{K}\sum_{y=1}^{K}\sum_{q=1}^{L} I(y = y_q^*)\frac{N_c(q)}{N(y)} \tag{2}$$

In the above two equations, I is indicator function. The choice of either should follow the considerations explained for the choice of average log likelihood.

Entropy. Similar to the case of average log likelihood, we give two options: overall entropy and mean class entropy. Again, the choice of either should follow the considerations explained for the choice of average log likelihood and correct classification probability. We define the *overall entropy* by

$$H(Q) = \frac{N_q}{N}\log\frac{N}{N_q} \tag{3}$$

We define *mean class entropy* by

$$H(Q) = \frac{1}{K}\sum_{y=1}^{K}\sum_{q=1}^{L} \frac{N_q(y)}{N(y)}\log\frac{N(y)}{N_q(y)} \tag{4}$$

Entropy has been used as a class impurity measure. But we use entropy as a measure of the consistence or generalization ability of the training results.

The performance measure function is defined by linearly combining $J(Q), H(Q)$ and $P_c(Q)$, as follows

$$T(Q) = W_J J(Q) + W_H H(Q) + W_c P_c(Q) \tag{5}$$

where W_J, W_H and W_c are weights. The choice of the weights depends on the pattern recognition task, normally being all non-negative. $T(Q)$ can always be written in an additive form $T(Q) = \sum_{q=1}^{L} T(q)$, where $T(q)$ is the contribution by an individual bin q. We define the performance measure of a sub-quantizer Q_r^u by $T(Q_r^u) = \sum_{q=r}^{u} T(q)$.

3. Quantization by dynamic programming

We are to find an L level quantizer Q optimizing $T(Q)$, also guaranteeing that the minimum number of data in each bin is $k \in \mathbb{N}$ and that identical data are put into the same bin regardless of their classes. We only put the decision boundaries in the middle of two neighboring input data. This affects the calculation of $J(Q)$, but it is trivial when sample size is not too small. This restriction prevents $J(Q)$ from overflow. When W_J is 0, the solution is indeed optimal since the exact placement of the decision boundaries are not important for the calculation of $H(Q)$ and $P_c(Q)$. A related problem is solved by dynamic programming, originally proposed in [3]. Their algorithm does not handle the minimum number points constraint and the identical data requirement. We also use a different quantizer performance measure with theirs.

We define the bin class density by $p(q|y) = \frac{N_q(y)/N(y)}{\Delta(q)}$. By definition of sub-quantizer measure, the additivity $T(Q_1^q) = T(Q_1^{q-1}) + T(q)$ gives rise to a dynamic programming algorithm. Assume $\mathcal{X}_N$ and $\mathcal{Y}_N$ are already sorted in non-decreasing order by x. Let $T[n,q]$ be the maximum performance measure from bin 1 to q when x_n is the largest data in bin q. Let $I[n,q]$ be the index to the smallest element in the q-th bin such that $T[n,q]$ is achieved. Let $T^1[i,n]$ be the quantizer measure contributed by a bin containing exactly x_i to x_n. The dynamic programming is described below.

Initialization. $T[0,0] = I[0,0] = 0$, $I[0,q] = -1$ for $q \in \{1,\cdots,L\}$, $I[n,0] = -1$ for $n \in \{1,\cdots,N\}$, $I[n,q] = -1$ for $(n,q) \in \{(n,q)|0 \le q < \max(1, n-(N-L))$ or $\min(n,L) < q \le L, 1 \le n \le N, 1 \le q \le L\}$.

Feasible decision boundary index set. The indices of the feasible data for being the smallest element in bin q form the set $\mathcal{A}_q = \{i|$, $i \le n-k+1$, $I[i-1,q-1] \ne -1$, $x_{i-1} \ne x_n$, $I[n,q] \ne -1$, $x_n \ne x_{n+1}\}$. $i \le n-k+1$ guarantees bin q contains at least k data. k plays a smoothing role, like the one in the nearest neighbor smoothing method. $I[i-1,q-1] \ne -1$ states that x_{i-1} must be feasible for the largest element in the previous bin $q-1$. $x_{i-1} \ne x_n$ enforces that the feasible largest element in the previous bin $q-1$ must not be the same as x_n, to avoid

splitting equal valued data into different bins. $x_n \neq x_{n+1}$ is not to split equal valued data. $I[n,q] \neq -1$ asserts that x_n must be feasible for the largest element of bin q.

Recurrence. If $\mathcal{A}_q$ is empty, then $I[n,q] \triangleq -1$, meaning x_n does not qualify for the largest element in bin q. Otherwise,

$$T[n,q] \triangleq \max_{i \in \mathcal{A}_q} T[i-1,q-1] + T^1[i,n] \quad (6)$$

$$I[n,q] \triangleq \operatorname*{argmax}_{i \in \mathcal{A}_q} T[i-1,q-1] + T^1[i,n] \quad (7)$$

We assert that $T[N,L]$ indeed achieves the maximum measure, the corresponding partition is an optimal solution, and the algorithm has time complexity $O(LN^2)$.

4. Density of a bin and smoothing

Weighted Averaging of Rounded Points (WARPing) is introduced in [5]. The data in each bin are rounded towards the center of the bin. Then the rounded data, instead of the original data set, are smoothed by a Parzen window method. For the density estimate of a single point, the time complexity is reduced from $O(N)$ to $O(L)$. We give a method that extends the notion of WARPing, which smoothes the quantized density using a generalized k nearest neighbor approach.

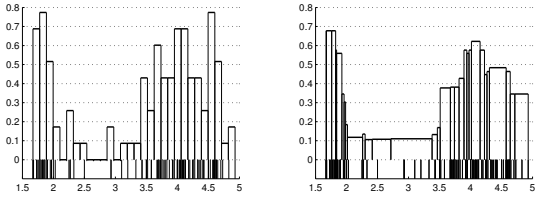
How to construct the k nearest neighborhood is an issue. Searching for the exact k-th nearest neighbor is not very pleasing because of the computation involved. We consider only neighboring bins instead of neighboring items of data, resulting a very fast algorithm for an approximate k nearest neighborhood.

Let $\Delta_k(q)$ be the width of a minimum neighborhood that contains at least k points. Let k_q be the actual number of points in the neighborhood. Then a smoothed probability density estimate of bin q is

$$p(q) = \frac{k_q}{\Delta_k(q) \sum_r \frac{k_r}{\Delta_k(r)} \Delta(r)} \quad (8)$$

5. Three examples

Old Faithful geyser eruption duration data. The data are a collection of durations, in minutes, of the Old Faithful geyser eruption [10]. The sample size is 107. The quantization level is 30. The optimally quantized histogram is shown in Fig. 1(b), where the quantizer performance measure is $T(Q) = J(Q)$ and number of neighbors $k = 18$. The equal bin width histogram, in Fig. 1(a), does not adjust its bin width to the data. The optimally quantized histogram allocates bin width by adapting to the data. For the low density region from about 2.1 to 3.4, six bins are used by the optimal quantization, while about fourteen bins are used by the equal bin width histogram. Around 2, 3.9, 4.3, 4.8, where the empirical



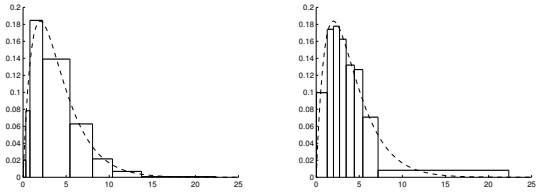
(a) Equal bin width histogram.

(b) Optimally quantized histogram.

Figure 1. Old Faithful geyser eruption duration (minutes) density estimates. Below each plot is the empirical density.

density changes rapidly, narrower bins are used by optimal quantization. This is mostly desirable because the narrower bins may lead to more accurate density estimates of the sharply changing density regions. The optimally quantized and smoothed density looks smoother, where three modes approximately at 1.7, 4.1 and 4.5 are identifiable, while the equal bin width histogram, not smoothed, is overly bumpy. This example shows the bin allocation efficiency of optimal histograms.

Chi-squared data. The data is simulated using Chi-squared distribution with 4 degrees of freedom. The sample size is 1000. The quantization level is 8. The density estimates (not smoothed) are shown in Fig. 2. The dashed line is the probability density function of the Chi-squared distribution. In Fig. 2(a), $T(Q) = J(Q)$. The



(a) Maximum average log likelihood quantization.

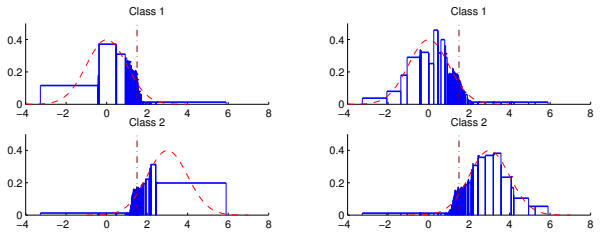
(b) Maximum entropy quantization.

Figure 2. Density estimates of Chi-squared data using optimal quantization.

bins are narrower for the region from 0 to 2 than for the region above 2. It is quite clear that the underlying density changes much more rapidly in $[0,2]$ than in $[2,\infty)$, corroborating the consistency result in [10]. In Fig. 2(b), $T(Q) = H(Q)$. The bins for the region around the mode at 2 are narrower than the region further away from the mode. The density of the region around the mode is larger than other regions. When entropy is max-

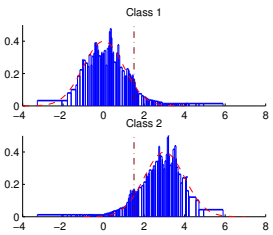
imized, each bin contains about the same number of data points. This naturally leads to narrower bins for regions of higher density and wider bins for regions of lower density. The rationale behind the entropy measure is that the least commitment should be made to the sample. This controls the generalization ability of the estimation result. On the other hand, the maximum likelihood histogram is always trying to find the best fit to the data and sometimes it may be overdone.

Two class normal data. Class 1 and 2 have 0 mean unit variance and 3 mean unit variance normal distributions, respectively. The sample sizes of both classes are 500. The quantization level is 100. The density estimates, using different performance measures and all smoothed with $k = 30$, are shown in Fig. 5. Solid and dashed lines represent the estimated and true densities, respectively. The dash-dotted lines are the decision boundary obtained by Bayesian rule with equal class prior and true densities. In Fig. 5(a), $T(Q) = P_c(Q)$. The



(a) $T(Q) = P_C(Q)$.

(b) $T(Q) = J(Q) + P_C(Q)$.



(c) $T(Q) = J(Q) + 0.5H(Q) + P_C(Q)$.

Figure 3. Optimal class histograms.

bins are consumed mostly around the boundary regions between the two classes. The regions far away from the class boundary have very wide bins and, therefore, inaccurate density estimates. But it suffices when the goal is classification. In Fig. 5(b) the performance measure is $T(Q) = J(Q) + P_c(Q)$. The class boundary region is still emphasized with more narrower bins. However densities of the far away regions from the class boundary take a much better shape, as compared to Fig. 5(a). We can see the density estimation around the class boundary still works well for classification. Fig. 5(c) shows the density estimates with $T(Q) = J(Q) + 0.5H(Q) + P_c(Q)$. Adding the entropy measure stabilizes the density estimates. It gives an even better fit towards the true densities. Still, the class boundary region is emphasized. Classification would work as well and an extra gain is that the underlying true densities are represented more accurately.

6. Conclusions

We described an approach to obtaining adaptive histograms by quantization and smoothing, optimizing a quantizer performance measure. The performance measure contains average log likelihood, entropy and correct classification probability measures. The optimal discretization is obtained with a dynamic programming algorithm. The density of each bin is estimated by a generalized k nearest neighbor algorithm. Experiments on different types of data are reported. The results show that optimal histograms are powerful in capturing the underlying true densities of the data using given resources. In addition, it is very efficient to use optimal histograms than other methods, e.g., the kernel density estimators, when dealing with massive amount of data.

References

[1] T. Elomaa and J. Rousu. General and efficient multisplitting of numerical attributes. *Machine Learning*, 36:201–44, 1999.

[2] U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Proc. 13th Int. Joint Conf. Artificial Intelligence*, pages 1022–1029, 1993.

[3] T. Fulton, S. Kasif, and S. L. Salzberg. Efficient algorithms for finding multi-way splits for decision trees. In *Proc. 12th Int. Conf. Machine Learning*, pages 244–251, 1995.

[4] R. M. Haralick. The table look-up rule. *Comm. in Stat. – Theory and Methods*, A5(12):1163–91, 1976.

[5] W. Härdle and D. W. Scott. Smoothing by weighted averaging of rounded points. *Comp. Stat.*, 7:97–128, 1992.

[6] L. B. Hearne and E. J. Wegman. Maximum entropy density estimation using random tessellations. In *Comp. Sci. & Stat.*, volume 24, pages 483–7, 1992.

[7] R. Kohavi and M. Sahami. Error-based and entropy-based discretization of continuous features. In *Proc. 2nd Int. Conf. KDD*, pages 114–119, 1996.

[8] G. Lugosi and A. B. Noble. Consistency of data-driven histogram methods for density estimation and classification. *Annals of Statistics*, 24:687–706, 1996.

[9] A. B. Nobel. Histogram regression estimation using data-dependent partitions. *Annals of Statistics*, 24(3):1084–1105, 1996.

[10] D. W. Scott. *Multivariate Density Estimation – Theory, Practice and Visualization*. John Wiley & Sons, 1992.

[11] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.