Query Selectivity Estimation Based on Improved V-optimal Histogram by Introducing Information about Distribution of Boundaries of Range Query Conditions

Dariusz Augustyn

Query selectivity estimation based on improved V-optimal histogram by introducing information about distribution of boundaries of range query conditions

Dariusz Rafal Augustyn

Silesian University of Technology, Institute of Informatics,
16 Akademicka St., 44-100 Gliwice, Poland
draugustyn@polsl.pl

Abstract. Selectivity estimation is a parameter used by a query optimizer for early estimation of the size of data that satisfies query condition. Selectivity is calculated using an estimator of distribution of attribute values of attribute involved in a processed query condition. Histograms built on attributes values from a database may be such representation of the distribution. The paper introduces a new query-distribution-aware V-optimal histogram which is useful in selectivity estimation for a range query. It takes into account either a 1-D distribution of attribute values or a 2-D distribution of boundaries of already processed queries. The advantages of qda-V-optimal histogram appears when it is applied for selectivity estimation of range query conditions that form so-called hot regions. To obtain the proposed error-optimal histogram we use dynamic programming method, Fuzzy C-Means clustering of a set of range boundaries.

Keywords: query selectivity estimation, data clustering, dynamic programming, distribution of range query conditions, V-optimal histogram

1 Introduction

Processing a database query by DBMS (Database Management System) consists of two phases – a prepare phase and an execute one. During the prepare phase the optimal query execution plan is obtained. This is done by a query cost optimizer which needs to estimate selectivity values of query selection conditions. For a simple single-table selection condition the selectivity is the number of rows satisfying the condition divided by the number of all table rows. For a simple single-table range condition based on x attribute with a continuous domain, it may be obtained as follows:

$$sel(Q(a < x < b) = \int_a^b f(x)dx. \tag{1}$$

where x – a table attribute, a and b – range query condition boundaries, $f(x)$ – a probability density function (PDF) of x attribute. Commonly, to obtain the selectivity value the query optimizer requires a histogram as a non-parametric estimator of PDF.

Since years there are well-known approaches to represent a 1-dimensional distribution of attribute values (e.g. based on histograms: equi-width, equi-height, max-diff, V-optimal [11] or spline representation). Some research concentrates on problem of obtaining a space-efficient representation of multi-dimensional representation (e.g. based on cosine series [13], wavelet transform [7], Bayesian network [9], self-tuning histograms [6, 12], kernel density estimator [10], and many other).

There are also some approaches which additionally use information about already processed queries ([8, 6, 12, 1, 2]). The proposed method uses it too, but it collects only information about the range conditions (values of boundaries), not about their real selectivity values obtained after a query execution (in opposite to the approaches presented in [8, 6]).

The proposed method (designated for building a representation of 1 dimensional distribution of attribute values) introduces a new type of histogram i.e. qda-V-optimal one (**q**uery-**d**istribution-**a**ware). It takes into account information about a range query workload. Such proposed hybrid representation uses either a 1-D distribution of attribute or a 2-D distribution of query conditions. We assume that DBMS should collect information about range boundaries of recently processed query conditions in a limited-length buffer. A result of clustering of boundaries values allows to modify some boundaries of V-optimal histogram buckets what tends to obtain a qda-V-optimal one.

This method may allow to create a better distribution representation than [1, 2], i.e. it is better adapted to a set of previously processed range queries. This results from taking into account full information about boundary pairs of processed query conditions (set of 2-D elements), in opposite to the approaches [1, 2] where we use only 1-D include function describing aggregated information about all ranges of processed queries. This implies a little greater storage requirement, of course.

The main contributions of the paper are as follows:

- the qda-V-optimal histogram – a representation of distribution of attribute values that it is partially adapted to a query workload, i.e. the representation resolution also depends on distribution of condition boundaries of recently processed range queries,
- the method of obtaining of a qda-V-optimal histogram, i.e. improving a V-optimal histogram by obtaining new histogram bucket's boundaries through clustering values of range boundaries,
- the method of reduction of complexity of the procedure by rejection weak clusters.

2 Motivating example – description of the proposed method

2.1 Exemplary attribute values distribution and its representation

To illustrate the concept of the proposed method of distribution representation we need a sample distribution of x attribute. We may use here any x distribution (although a non-uniform one is rather expected).

In the example we assume the one which is based on superposition of $G = 4$ truncated Gaussian clusters with bounded support (limited to [0 1]), where parameters of used univariate normal distributions are shown in Tab. 1. PDF is defined as follows:

$$f(x) = \sum_{i=1}^{G} p_i \, \mathrm{PDF}_{\mathrm{TN}}(x, m_i, \sigma_i, 0, 1), \tag{2}$$

where $\mathrm{PDF}_{\mathrm{TN}}(x, m_i, \sigma_i, l, r)$ is PDF of truncated normal distribution with a support based on interval $[l\ r]$:

$$\mathrm{PDF}_{\mathrm{TN}}(x, m_i, \sigma_i, l, r) = \frac{\frac{1}{\sigma_i}\phi(\frac{x-m_i}{\sigma_i})}{\Phi(\frac{r-m_i}{\sigma_i}) - \Phi(\frac{l-m_i}{\sigma_i})}, \tag{3}$$

where ϕ is PDF of $N(0,\ 1)$ and Φ is cumulative density function (CDF) of $N(0,\ 1)$.

Table 1. Parameters of clusters used in the definition of exemplary PDF of x attribute

i – cluster number	1	2	3	2
p_i	0.25	0.25	0.3	0.2
m_i	0.2	0.8	0.6	0.7
σ_i	0.1	0.1	0.01	0.01

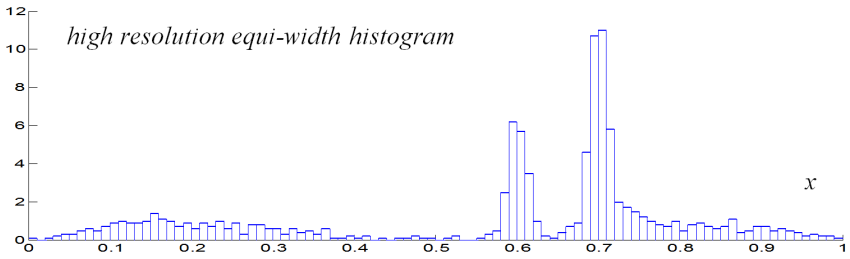

Fig. 1. High resolution representation of x attribute distribution – the equi-width histogram.

The distribution consists of two narrow clusters ($i = 3, 4$; with relatively small σ_i) and two wide ones ($i = 1, 2$; with big σ_i). The high resolution empirical histogram with 100 intervals (built on 1000 samples) describing $f(x)$ is shown in Fig. 1.

Let us consider to use a V-optimal histogram [11] as a representation of the distribution. $B = 10$ is the assumed number of buckets in this histogram. Using the method of dynamic programming [11] we obtained such distribution of bucket boundaries (see Fig. 2) denoted by $b_{vo\,j}$ for $j = 1, \ldots, B+1$ (where $b_{vo\,1} = 0$ and $b_{vo\,B+1} = b_{vo\,11} = 1$) that sum of variances of frequencies in buckets is the smallest. The frequencies are taken from the equi-width histogram (Fig. 1) with a relatively high resolution (with $10 \cdot B$ buckets). The result domain division is presented in Fig. 2 (dashed vertical lines). The result V-optimal histogram is shown in Fig. 3.

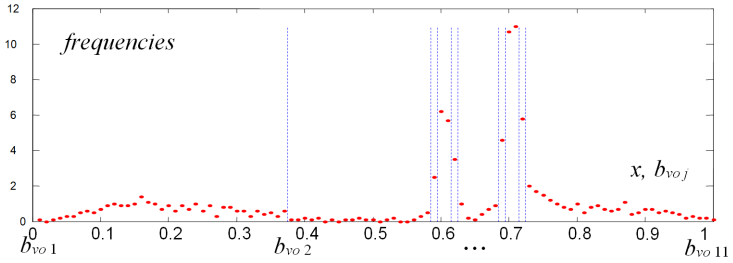

Fig. 2. $(b_{vo\,j})$ for $j = 1, \ldots, B+1$ – found boundaries of V-optimal histogram.

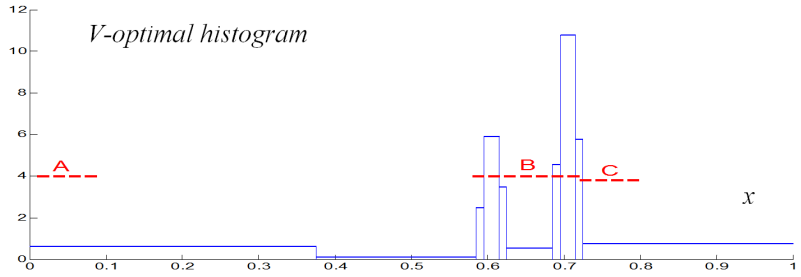

Fig. 3. The V-optimal histogram with $B = 10$ buckets – the low resolution representation of attribute value distribution. Hot regions (A, B, C) describing the range query interval distribution.

2.2 Exemplary of distribution of range query condition bounds

In the proposed method we assume that we collect conditions of most recently executed queries.

Let us assume the situation from real world where a set of processed query conditions forms so-called hot regions.

In our example we used a superposition of distributions of hot regions A, B, C (Fig. 3), and the truncated 2D-uniform distribution.

Region **A** defines query boundaries based on the interval $[\boldsymbol{a}, \boldsymbol{b}] = [\mathbf{0.01}, \mathbf{0.09}]$ i.e. we have a values located near 0.01 (and $0 \leq a \leq 1$) and we have b values located near 0.09 (and $a \leq b \leq 1$). Here, we assumed the following density functions for either left endpoints of intervals belonging to hot region A or right ones: $\mathrm{PDF}_{a\mathrm{A}}(a) = \mathrm{PDF}_{\mathrm{TN}}(a, m_{a\mathrm{A}}, \sigma_{a\mathrm{A}}, 0, 1) = \mathrm{PDF}_{\mathrm{TN}}(a, 0.01, 0.0005, 0, 1)$ and $\mathrm{PDF}_{b\mathrm{A}}(b|a) = \mathrm{PDF}_{\mathrm{TN}}(b, m_{b\mathrm{A}}, \sigma_{b\mathrm{A}}, a, 1) = \mathrm{PDF}_{\mathrm{TN}}(b, 0.09, 0.0001, a, 1)$. Those relatively small values of $\sigma_{a\mathrm{A}}$ and $\sigma_{b\mathrm{A}}$ (comparing to values of $m_{a\mathrm{A}}$ and $m_{b\mathrm{A}}$) cause that all query ranges belonging to region A are almost the same. Thus, pairs (a, b) that belong to region A are described by:

$$\begin{aligned} &\mathrm{PDF}_{\mathrm{A}}(a, b) = \mathrm{PDF}_{a\mathrm{A}}(a) \cdot \mathrm{PDF}_{b\mathrm{A}}(b|a) = \\ &= \mathrm{PDF}_{\mathrm{TN}}(a, m_{a\mathrm{A}}, \sigma_{a\mathrm{A}}, 0, 1) \cdot \mathrm{PDF}_{\mathrm{TN}}(b, m_{b\mathrm{A}}, \sigma_{b\mathrm{A}}, a, 1) = \\ &= \mathrm{PDF}_{\mathrm{TN}}(a, 0.01, 0.0005, 0, 1) \cdot \mathrm{PDF}_{\mathrm{TN}}(b, 0.09, 0.0001, a, 1). \end{aligned} \tag{4}$$

Analogously, we define regions B and C. Region **B** defines query boundaries based on the interval $[\boldsymbol{a}, \boldsymbol{b}] = [\mathbf{0.58}, \mathbf{0.72}]$ i.e.:
$\mathrm{PDF}_{a\mathrm{B}}(a) = \mathrm{PDF}_{\mathrm{TN}}(a, m_{a\mathrm{B}}, \sigma_{a\mathrm{B}}, 0, 1) = \mathrm{PDF}_{\mathrm{TN}}(a, 0.58, 0.0005, 0, 1)$ and
$\mathrm{PDF}_{b\mathrm{B}}(b|a) = \mathrm{PDF}_{\mathrm{TN}}(b, m_{b\mathrm{B}}, \sigma_{b\mathrm{B}}, a, 1) = \mathrm{PDF}_{\mathrm{TN}}(b, 0.72, 0.0001, a, 1)$. Thus
$\mathrm{PDF}_{\mathrm{B}}(a, b) = \mathrm{PDF}_{a\mathrm{B}}(a) \cdot \mathrm{PDF}_{b\mathrm{B}}(b|a)$.
Region **C** defines query boundaries based on the interval $[\boldsymbol{a}, \boldsymbol{b}] = [\mathbf{0.72}, \mathbf{0.8}]$ i.e.: $\mathrm{PDF}_{a\mathrm{C}}(a) = \mathrm{PDF}_{\mathrm{TN}}(a, m_{a\mathrm{C}}, \sigma_{a\mathrm{C}}, 0, 1) = \mathrm{PDF}_{\mathrm{TN}}(a, 0.72, 0.0005, 0, 1)$ and
$\mathrm{PDF}_{b\mathrm{C}}(b|a) = \mathrm{PDF}_{\mathrm{TN}}(b, m_{b\mathrm{C}}, \sigma_{b\mathrm{C}}, a, 1) = \mathrm{PDF}_{\mathrm{TN}}(b, 0.8, 0.0001, a, 1)$. Thus
$\mathrm{PDF}_{\mathrm{C}}(a, b) = \mathrm{PDF}_{a\mathrm{C}}(a) \cdot \mathrm{PDF}_{b\mathrm{C}}(b|a)$.

PDF of truncated 2D-unifrom distribution described as follows:

$$\mathrm{PDF}_{\mathrm{T2D-uniform}}(a, b) = \begin{cases} 2 & \text{for } 0 \leq a \leq 1 \wedge 0 \leq b \leq 1 \wedge a \leq b \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

is presented in Fig. 4b. The truncated 2D-uniform was used for introducing (into a result boundaries distribution) some events that not all pairs (a, b) come from hot regions, i.e. some of them are uniformly distributed in $[0\ 1]^2$ space (subject to $b \geq a$).

Finally the result distribution of query bounds is assumed as follows:

$$\begin{aligned} &\mathrm{PDF}_{ab}(a, b) = 0.3 \cdot \mathrm{PDF}_{\mathrm{A}}(a, b) + 0.3 \cdot \mathrm{PDF}_{\mathrm{B}}(a, b) + \\ &+0.2 \cdot \mathrm{PDF}_{\mathrm{C}}(a, b) + 0.2 \cdot \mathrm{PDF}_{\mathrm{T2D-uniform}}(a, b). \end{aligned} \tag{6}$$

$M = 20$ samples were generated according to the distribution given by eq. (6). The generated pairs (a_j, b_j) for $j = 1, \ldots, M$ are shown in Fig. 4a. They form a set named *Qset*. *Qset* is a collection of query boundaries from last M queries.

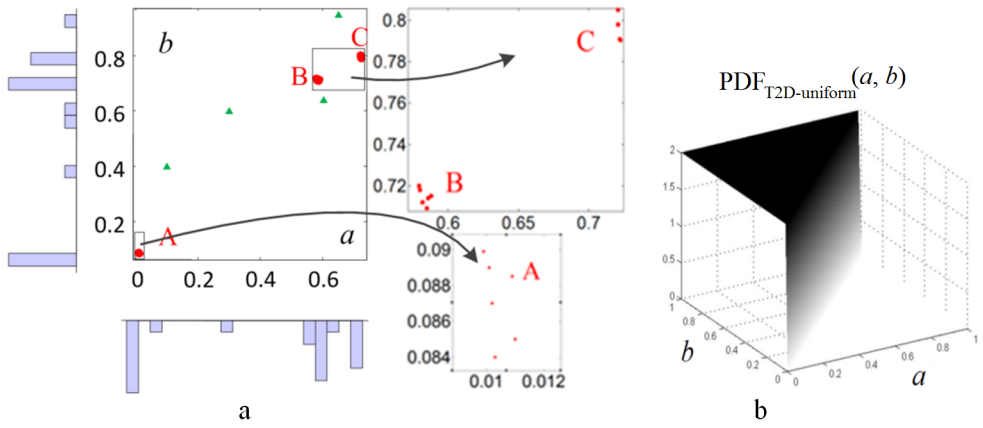

Fig. 4. a) *Qset* – the set of sample query bounds – the exemplary set of pairs (a_j, b_j) for $j = 1, \ldots, M (M = 20)$, generated according to the $\mathrm{PDF}_{ab}(a, b)$. b) Bivariate PDF of the truncated 2D uniform distribution (eq. (5)).

Those pairs that come from truncated 2-D uniform distribution are represented by 4 triangles in Fig. 4a.

In our method we recommend rather similar values of M and B, i.e. they should have the same order of magnitude (the size of metadata needed for describing range conditions should be rather similar to the size of metadata describing an attribute values distribution).

2.3 Verifying the applicability of the proposed method

A verification step is performed before any essential activities of the proposed method. We create the 1-dimensional vector S which contains either a_j or b_j elements. S presented in Fig. 5, was built on all pairs (a_j, b_j) from *Qset* (i.e. for $j = 1, \ldots, M$). In our example S consists of $2 \cdot M = 40$ elements.

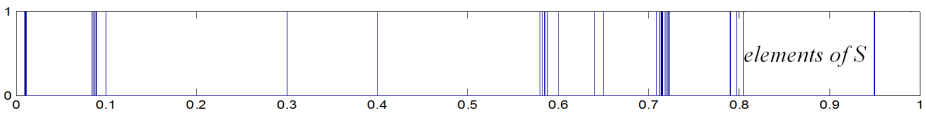

Fig. 5. Distribution of values from S which consists of either a or b values from *Qset*.

In this step of the proposed method, we check if elements of S do not represent the 1-D uniform distribution based on [0 1] interval. This is done by some well-known methods of nonparametric test for the equality of an empirical distribution and a reference one (e.g. we use chi-square test or Kolmogorov-Smirnov

one). This is done to reject (at the assumed significance level) the null hypothesis that the data in S comes from the [0 1] uniform distribution. As we may expect, the exemplary set S shown in Fig. 5 is not a sample set coming from the uniform distribution (at 5% significance level). If S represented data from the uniform distribution then there would be no hot regions in $Qset$, so there would be no reason to improve a classic V-optimal histogram.

2.4 Using K-fold cross validation for obtaining the optimal clusters of boundaries of range query condition

The proposed method introduces clustering of query bounds. The obtained set of centers of clusters will determine some new boundaries of buckets in a histogram which represents x distribution. Such histogram based on a V-optimal one will be named qda-V-optimal (query-distribution-aware).

To obtain the optimal number of clusters, denoted by C_{opt}, we use well-known K-fold cross validation method. To obtain C_{opt} we minimize a score function which is based on a mean relative selectivity error (described further by eq. (9) and (10)).

In our example, we assumed 5-fold cross validation. We divide $Qset$ into learning sets $Qset\_learn_k$ and test sets $Qset\_test_k$. There are $K = 5$ different either learning sets or test ones. Each $Qset\_learn_k$ has $(K-1)/K \cdot M = 16$ pairs of a and b. Each $Qset\_test_k$ has $1/K \cdot M = 4$ pairs. All sets $Qset\_learn_k$ and $Qset\_test_k$ satisfy the following constraint:

$$\forall_{k=1,\ldots,K} \;\; Qset = Qset\_learn_k \cup Qset\_test_k. \tag{7}$$

To find the optimal C_{opt} by applying K-fold cross validation we use the score function values calculated for $Qset\_test_k$ (for $k = 1, \ldots, K$).

Let us define S_k ($S_k \subset S$) as a one-dimensional vector which contains either a values and b ones from $Qset\_learn_k$.

To obtain clusters in S_k we use the very well known method – FCM (Fuzzy C-Means) [4].

Clusters from S_k (which is based on $Qset\_learn_k$) will be used for finding a new set of bucket boundaries and construct a qda-V-optimal. The relevant $Qset\_test_k$ will be used for validate accuracy of this qda-V-optimal histogram the by calculations of selectivity estimation errors.

Steps of the method will be illustrated using S_1 – an exemplary S_k. Elements of vector S_1 are shown in Fig. 7 (values of S_1 are presented using vertical dashed lines).

2.5 Clustering query boundary values from learn set and rejecting either low cardinality clusters or wide ones

We assumed that the support of $f(x)$ is limited, i.e. $\min(x) = 0$ and $\max(x) = 1$ (Fig. 1), so this determines the first boundary of qda-V-optimal histogram and the last one. Thus, during the optimization we may only change positions of

maximally $B-1$ internal bucket boundaries of qda-V-optimal histogram. This determines the maximum of the number of clusters. Thus, C – the number of clusters – may vary, but it has to be less than the number of internal buckets in a standard V-optimal histogram, i.e. $2 \leq C \leq B-1$.

We propose some constraints for cluster's properties. This allows to eliminate so-called weak clusters. We want to consider only:

- enough narrow clusters, i.e. such ones that 2· standard deviation of cluster element values is less than some assumed value – ϵ. In the example we assume $\epsilon = 1/B = 1/10$ (we use here $1/B$ – length of an equi-width histogram with B buckets and with domain [0 1]),
- only high cardinality clusters, i.e. such ones that the number of elements in a cluster is greater than $N_{min\_elem\_in\_cluster}$ – some assumed fraction of all S elements. In the example we assume $N_{min\_elem\_in\_cluster} = 10\%$ of $M = 0.1 \cdot 20 = 2$.

For any given C we may find C_{acc} – the number of accepted clusters ($C_{acc} \leq C$) – that satisfy the mentioned-above criteria.

In Fig. 6a, we present the result of clustering of the exemplary vector S_1, which elements come from $Qset\_learn_1$. We consider $C = 2, \ldots, B-1 = 2, \ldots, 9$. In Fig. 6a we can see the properties of accepted clusters:

- ***me*** – center (median) value of an accepted cluster,
- ***width*** – width of a cluster (presented as a vertical interval with whiskers) with length given by 2· std of the cluster elements values,
- ***card*** – cardinality of a cluster.

According to the assumed constraints given by $(\epsilon, N_{min\_elem\_in\_cluster})) = (0.1, 2)$ we obtain the relevant numbers of accepted clusters as follows $C_{acc} = 1, 3, 3, 4, 4, 4, 4$ for given $C = 3, 4, \ldots, 9$ (Fig. 6a, 6b). For $C = 2$ there were no clusters that satisfy constraints for cluster's properties.

For decreasing values of C, cardinalities of clusters rather increase but the clusters also become wider. Dependency between $total\_card$ – the total sum of cardinalities of accepted clusters and C is presented in Fig. 6c. Dependency between $mean\_width$ – the mean width of accepted clusters and C is shown in Fig. 6d.

The described heuristic procedure of rejecting weak clusters allows to limit the number of considered C values. For small or big values of C there will be no accepted clusters, i.e. the relevant C_{acc} will equal 0 (too wide clusters or too low cardinality ones). In further steps we will only consider those C that satisfy $C_{acc}(C) > 0$. This advantage of the cluster rejection may be useful for high value of B (and high value of $C = 2, \ldots, B-1$, respectively) because in real DBMSes we may expect such values of B (i.e. commonly hundreds).

2.6 Creating qda-V-optimal histogram – rejecting boundaries of V-optimal histogram that are close to centers of accepted clusters

Let us deeply consider the case of S_1 clustering where the number of clusters $C = 6$ (so the relevant C_{acc} equals 4 (Fig. 6a, 6b)).

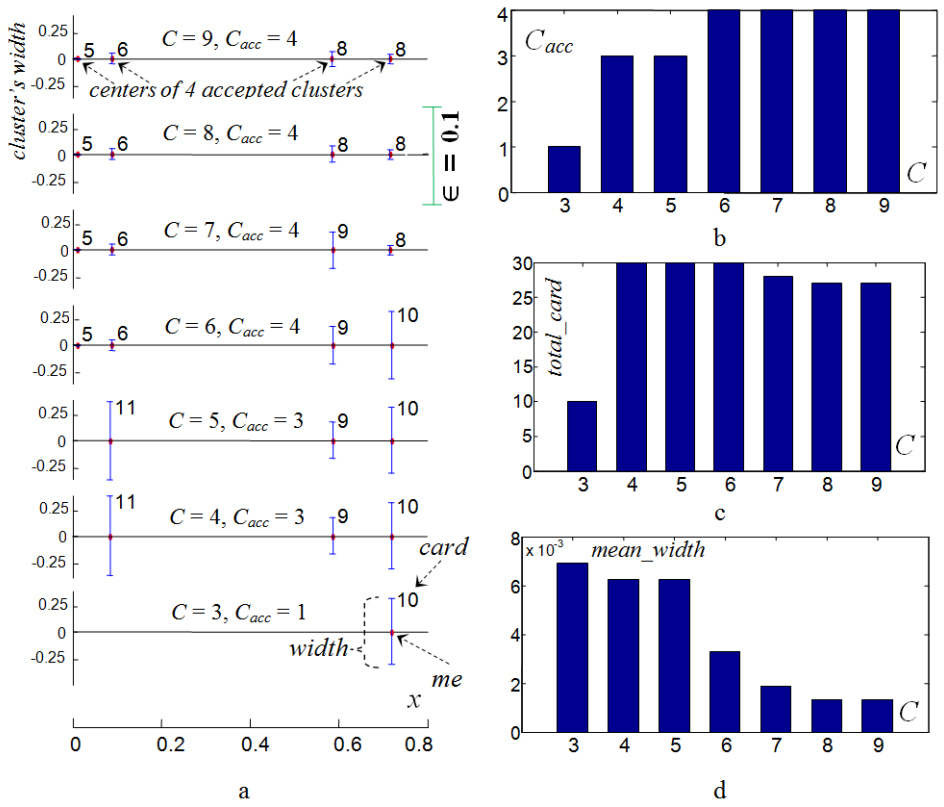

Fig. 6. a) Sets of accepted clusters for different values of C (the number of clusters used in FCM method). Dependency between C and: b) the number of accepted clusters, c) the total sum of cardinality of accepted clusters, d) the mean width of accepted clusters.

$C_{acc} = 4$ accepted centers of clusters (Fig. 7) become some of bounds of the required qda-V-optimal.

The rest of $B + 1 - C_{acc} = 11 - 4 = 7$ boundaries of the qda-V-optimal histogram will be selected from the $\{b_{voj}\}$ i.e. boundaries of the V-optimal histogram (Fig. 2 and 3). We want to reject those that are placed near the centers of accepted clusters. To do this we may find the distance between a selected b_{voj} and the set centers of accepted clusters as follows:

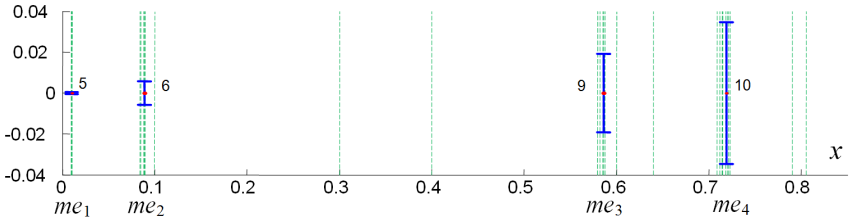

Fig. 7. Four accepted cluster defined by the series of medians: $me_1, \ldots, me_{Cacc} = 0.0103, 0.08875, 0.586, 0.71895$. Dashed lines present 32 element values of S_1.

$$d_j = \min_{i=1,\ldots,Cacc} |b_{vo\,j} - me_i|. \tag{8}$$

Finally, we may find 7 boundaries from all $b_{vo\,j}$ $(j = 1, \ldots, B+1)$ that have the largest d_j. So, we have eliminated those 4 bounds: $(b_{vo\,3}, b_{vo\,4}, b_{vo\,9}, b_{vo\,10}) = (0.585, 0.595, 0.715, 0.725)$ that are not enough distant from the centers of accepted clusters $(me_1, \ldots, me_4)$.

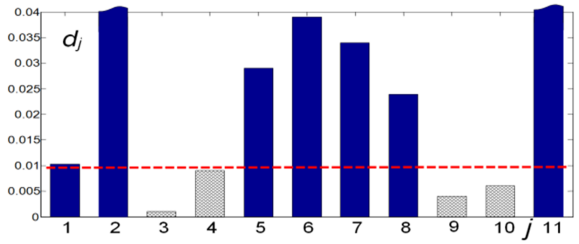

Fig. 8. Selecting 7 top long distance boundaries (solid rectangles with values above the dashed horizontal line) from all boundaries of V-optimal histogram.

This allows to obtain the all boundaries of the result qda-V-optimal histogram $(b_{vo\,1}, \ldots, b_{vo\,11})$ that are based either on centers of accepted clusters $(me_1, \ldots, me_4)$ or selected boundaries from the V-optimal histogram, i.e. $b_{vo\,j}$ where $j = 1, 2, 5, 6, 7, 8, 11$.

Using frequencies from the high resolution equi-width histogram (Fig. 1) we may find values of the result qda-V-optimal histogram (Fig. 9a) which buckets are based on found boundaries $(b_{vo\,1}, \ldots, b_{vo\,11})$. This histogram will be scored by the mean relative selectivity estimation error in a next step of the proposed method.

Along $[0.01, 0.09]$ (hot region A) the qda-V-optimal histogram (Fig. 9) has 3 buckets while the standard V-optimal one (Fig. 3, 9b) has only one bucket. Thus, this qda-V-optimal histogram is better adapted to the query conditions belonging to the hot region A. (This improvement is at the expense of losing the

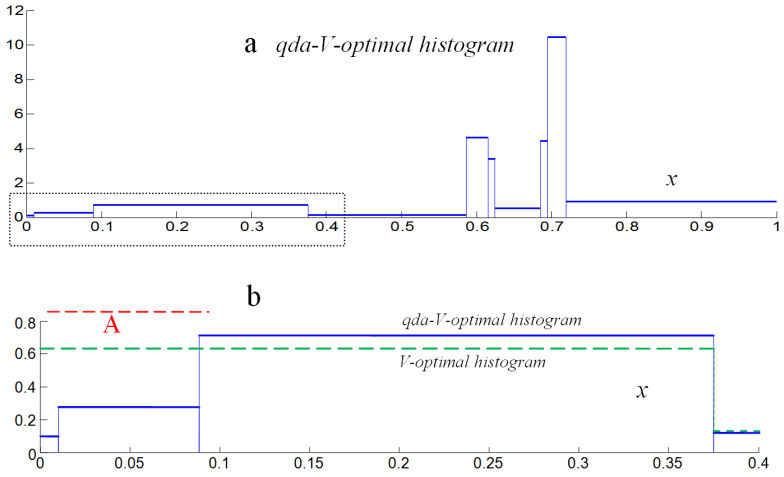

Fig. 9. a) The qda-V-optimal histogram (based on V-optimal one) built by clustering the values form S_1 (based on $Qset\_learn_1$) for the assumed number of cluster – $C = 6$ (and the relevant number of accepted clusters – $C_{acc} = 4$).
b) Zoomed part of the qda-V-optimal histogram (solid line) and the relevant V-optimal one (dashed line) for $x \in [0, 0.4]$. There are 3 buckets of the qda-Voptimal histogram in $x \in [0.01, 0.09]$ (hot region A).

accuracy of distribution representation in intervals not overlapped by hot query range regions).

2.7 Obtaining the optimal number of clusters of query boundaries – minimizing mean error of selectivity estimations based on qda-V-optimal histogram

Let us introduce the relative selectivity error as follows:

$$RelSelErr_H(a,b) = RelSelErr_H(Q(a < x < b)) = \frac{|\widehat{sel}_H(Q) - sel(Q)|}{sel(Q)} \cdot 100\%. \tag{9}$$

$\widehat{sel}_H$ denotes an approximated selectivity value based on some histogram H, where H is VO (V-optimal histogram) or $qdaVO$ (qda-V-optimal one). sel is an exact selectivity value (here calculated using a high resolution equi-width histogram, like this from Fig. 1).

Mean relative error of selectivity for some $Qset$ (some set of query conditions) and some histogram H is given as follows:

$$MeanRelSelErr_H(Q) = \text{maean}_{(a,b)\in Qset}\ \ RelSelErr_H(a,b). \tag{10}$$

Using the V-optimal histogram (Fig. 3) and the test set of query conditions $Qset\_test_1$, we obtained $MeanRelSelErr_{VO}(Qset\_test_1) \approx 38.3\%$. Using the

qda-V-optimal histogram (Fig. 9a) and $Qset\_test_1$, we obtained $MeanRelSelErr_{qdaVO}(Qset\_test_1) \approx 11.9\%$ which is a better result than that obtained from the V-optimal histogram.

Finally, using the mentioned K-fold validation method, we may obtain C_{opt}, i.e. the error-optimal number of clusters, which gives the smallest mean value of $MeanRelErrSel_{qdaVO}$ over all $Qset\_learn_k$ and $Qset\_test_k$ (for $k = 1, \ldots, K$).

For the assumed $Qset$ (Fig. 4a), using 5-fold cross validation, we obtained $C_{opt} = 6$ (Fig. 10). Mean relative errors for V-optimal (obtained by using the same sets $Qset\_test_k$) are also presented in Fig. 10. For given distribution of attribute x (Fig. 1) and set of query condition defined by $Qset$ (Fig. 4), applying the error-optimal qda-V-optimal histogram allows to obtain smaller mean relative selectivity estimation error (11.1%) than applying the V-optimal histogram (32.2%).

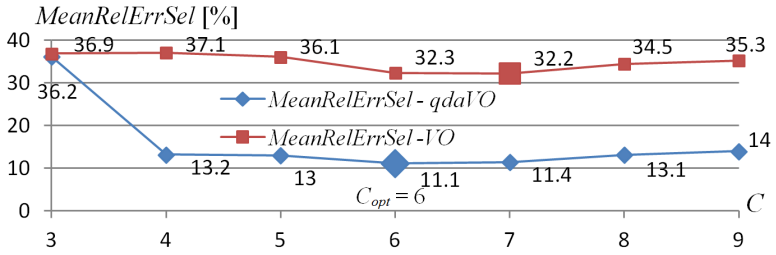

Fig. 10. K-fold cross validating the method of selectivity estimation using qda-V-optimal histogram for query condition given by $Qset$. Obtaining the optimal number of clusters ($C_{opt} = 6$) with the smallest mean relative selectivity ($\approx 11.1\%$)

For the exemplary distributions given in subsections 2.1 and 2.2, the qda-V-optimal histogram seems significantly better than the relevant V-optimal one.

3 The algorithm of the proposed method

The proposed method is designated to obtain a qda-V-optimal histogram which may be applied for range query selectivity estimation.

The main goal is to find the error-optimal histogram for assumed B – an arbitrary given number of histogram buckets. To build a qda-V-optimal histogram we assume that we already have collected elements of $Qset$, i.e. M last processed range query conditions that operated on x attribute.

The proposed method should be invoked during execution of so-called update statistics command for x attribute.

The proposed method consists of the following steps:

1. Create a high resolution equi-width histogram (this is a temporary structure which is removed after finishing the method; the number of buckets of the equi-width histogram should be significantly greater than B).

2. Create a classic V-optimal histogram with B buckets using the equi-high histogram.
3. Check applicability of the proposed method, i.e. verify that elements from S (a set of values built on elements of $Qset$) do not come from uniform distribution (if so, than there are no hot regions of query conditions, so only the V-optimal histogram is returned as a result and the method is finished).
4. Using K-fold cross validation, obtain C_{opt} – the error-optimal number of clusters of query boundaries which determines the qda-V-optimal histogram (the equi-histogram histogram and the V-optimal one are used; elements form subset of S are clustered using FCM method; the constraints (small width or low cardinality) are applied for elimination weak clusters; those boundaries of the V-optimal histogram are used in the qda-V-optimal one, that they are enough distant from centers of accepted clusters).
5. Verify the superiority of the qda-V-optimal histogram over the V-optimal one for given $Qset$ (verifying that the qda-V-optimal histogram based on C_{opt} allows to calculate query selectivities for $Qset$ elements with less mean relative error than the V-optimal one; if so, the qda-V-optimal histogram is returned else the V-optimal one is).

4 Conclusions

The paper describes the method of query selectivity estimation based on improved V-optimal histograms. The proposed method additionally takes into account information about the query conditions of previously processed range queries. A proposed new qda-V-optimal histogram may allow to estimate the selectivities with small mean relative errors, especially when query conditions form so-called hot regions.

In future we may consider to introduce importance weights for conditions from the buffer (a newer condition is more important). We may also consider of applying a different policy of discarding elements from the full buffer (e.g. not to use LRU method (least recently used) but use hit ratios and LFU method (least frequently used)).

Further work may concentrate on a problem when a future query workload is not well represented by boundary samples already collected in a buffer, i.e. in last M query conditions. Some simple way to solve overlearning a qda-V-optimal histogram by no enough representative $Qset$ is to introduce some additional random samples (according to the truncated 2D-uniform distribution) to decrease too a large impact of a query distribution on the qda-V-optimal histogram construction.

The proposed method of obtaining a qda-V-optimal should be invoked during update statistics (not during on-line query processing) and it is not so time critical operation like a selectivity estimation. However, because of the complexity of the proposed algorithm, we consider using capabilities of GPGPU (General-Purpose computing on Graphics Processor Units) having in mind that parallel efficient GPU becomes useful in database processing (e.g. [5, 3]).

References

1. Augustyn, D.R.: Query-condition-aware histograms in selectivity estimation method. In: Man-Machine Interactions 2, *Advances in Intelligent and Soft Computing*, vol. 103, pp. 437–446. Springer Berlin Heidelberg (2011). DOI 10.1007/978-3-642-23169-8_47. URL http://dx.doi.org/10.1007/978-3-642-23169-8_47
2. Augustyn, D.R.: Query-condition-aware v-optimal histogram in range query selectivity estimation. Bulletin of the Polish Academy of Sciences. Technical Sciences **62**(2), 287–303 (2014). DOI 10.2478/bpasts-2014-0029. URL http://dx.doi.org/10.2478/bpasts-2014-0029
3. Augustyn, D.R., Zederowski, S.: Applying cuda technology in dct-based method of query selectivity estimation. In: New Trends in Databases and Information Systems, pp. 3–12. Springer (2013)
4. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell, MA, USA (1981)
5. Breß, S., Beier, F., Rauhe, H., Sattler, K.U., Schallehn, E., Saake, G.: Efficient co-processor utilization in database query processing. Inf. Syst. **38**(8), 1084–1096 (2013). DOI 10.1016/j.is.2013.05.004. URL http://dx.doi.org/10.1016/j.is.2013.05.004
6. Bruno, N., Chaudhuri, S., Gravano, L.: Stholes: A multidimensional workload-aware histogram. SIGMOD Rec. **30**(2), 211–222 (2001). DOI 10.1145/376284.375686. URL http://doi.acm.org/10.1145/376284.375686
7. Chakrabarti, K., Garofalakis, M., Rastogi, R., Shim, K.: Approximate query processing using wavelets. The VLDB Journal **10**(2-3), 199–223 (2001). URL http://dl.acm.org/citation.cfm?id=767141.767147
8. Chen, C.M., Roussopoulos, N.: Adaptive selectivity estimation using query feedback. SIGMOD Rec. **23**(2), 161–172 (1994). DOI 10.1145/191843.191874. URL http://doi.acm.org/10.1145/191843.191874
9. Getoor, L., Taskar, B., Koller, D.: Selectivity estimation using probabilistic models. SIGMOD Rec. **30**(2), 461–472 (2001). DOI 10.1145/376284.375727. URL http://doi.acm.org/10.1145/376284.375727
10. Gunopulos, D., Kollios, G., Tsotras, J., Domeniconi, C.: Selectivity estimators for multidimensional range queries over real attributes. The VLDB Journal **14**(2), 137–154 (2005). DOI 10.1007/s00778-003-0090-4. URL http://dx.doi.org/10.1007/s00778-003-0090-4
11. Jagadish, H.V., Poosala, V., Koudas, N., Sevcik, K., Muthukrishnan, S., Suel, T.: Optimal histograms with quality guarantees. In: In VLDB, pp. 275–286 (1998)
12. Khachatryan, A., Müller, E., Böhm, K., Kopper, J.: Efficient selectivity estimation by histogram construction based on subspace clustering. In: Proceedings of the 23rd international conference on Scientific and statistical database management, SSDBM'11, pp. 351–368. Springer-Verlag, Berlin, Heidelberg (2011). URL http://dl.acm.org/citation.cfm?id=2032397.2032426
13. Yan, F., Hou, W.C., Jiang, Z., Luo, C., Zhu, Q.: Selectivity estimation of range queries based on data density approximation via cosine series. Data Knowl. Eng. **63**(3), 855–878 (2007). DOI 10.1016/j.datak.2007.05.003. URL http://dx.doi.org/10.1016/j.datak.2007.05.003