Query Workload Aware Multi-histogram Based on Equi-width Sub-histograms for Selectivity Estimations of Range Queries

Dariusz Rafał Augustyn(✉)

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

Abstract. Query optimizer uses a selectivity parameter for estimating the size of data that satisfies a query condition. Selectivity value calculations are based on some representation of attribute values distribution e.g. a histogram. In the paper we propose a query workload aware multi-histogram which contains a set of equi-width sub-histograms. The multi-histogram is designated for single-attribute-based range query selectivity estimating. Its structure is adapted to a 2-dimensional distribution of conditions of last recently processed range queries. The structure is obtained by clustering values of boundaries of query ranges. Sub-histograms' resolutions are adapted to a variability of a 1-dimensional distribution of attribute values.

Keywords: Selectivity estimation · Range query · Multi-histogram · embedded sub-histogram · Query workload · Data clustering · Bucket boundaries distribution · Variability metrics

1 Introduction

Selectivity factor is used by a database query optimizer to choose the best query execution plan. It is needed for an early estimation of size of the data that satisfying a query condition. 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 range condition based on single attribute x with continuous domain, it may be defined as follows:

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

where x – a table attribute, a and b – range boundaries, $f(x)$ – a probability density function (PDF) that describes x attribute values distribution.

There are many approaches to representing an attribute values distribution using different types of histogram [8]. Most of them use only an information

S. Kozielski et al. (Eds.): BDAS 2015, CCIS 521, pp. 47–59, 2015.
DOI: 10.1007/978-3-319-18422-7_4

about x values distribution but also there are some that take into account an information about query workload [6, 5, 9, 11, 1–3].

The proposed method also uses information about processed queries, but it only collects data about the range conditions (values of range boundaries), not about their real selectivity values obtained just after a query execution (like the approaches presented in [6, 5, 9, 11]. [7]. Some of those approaches to query-workload-based selectivity estimation (so-called feedback driven) are dedicated for multi-dimensional queries (m-D range queries with conditions based on many attributes), e.g. the approaches that use: self-tunning histogram and STHoles [5, 9, 10], ISOMER – the maximum entropy based algorithm for feedback-driven m-D histogram creation [11], proactive and reactive m-D histogram [7].

In this paper we introduce a new representation of attribute values distribution – a multi-histogram – which consists on non-overlapping equi-width sub-histograms. Domains of sub-histograms depend on 2-D distribution of pairs (a, b) that describe range boundaries of last recently processed queries. Such 2-D representation is more detailed than 1-D one given by the including function proposed in [1, 2]. In the proposed approach we use clustering of query range boundary values (like in [3]) for adapting the multi-histogram to historical data about conditions of processed range queries. This allows to divide whole domain of multi-histogram and to use simple equi-width histograms (called sub-histograms) in obtained sub-domains (there is no usage of sub-histogram in [3]).

The contributions of the paper are as follows:

- query workload aware multi-histogram representation of an attribute values distribution,
- methods of improvement of sub-histograms' resolutions (also partially adapted to query workload) by increasing them in domain regions of high variability of PDF(x).

2 Description of the Proposed Method and the Example of Usage

2.1 Exemplary Attribute Values Distribution

The proposed method will be presented by using a sample distribution of x attribute [3]. To build an exemplary distribution representation we use a pseudorandom generator based on superposition of $G = 4$ Gaussian clusters with bounded support (limited to $[0, 1]$), where parameters of used univariate truncated normal distributions are shown in table 1. The relevant PDF is defined as follows:

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

The distribution consists of two narrow clusters (no 3, 4 with small sigmas) and two wide ones (no 1, 2). Of course, we may use here any type of 1-D distribution, based not only on Gaussian clusters.

Table 1. Parameters of clusters used in the definition of exemplary PDF of x attribute [3]

$Cluster\ no$	1	2	3	4
p_i	0.25	0.25	0.3	0.2
m_i	0.2	0.8	0.6	0.8
σ_i	0.12	0.12	0.001	0.01

In the further consideration we will use a high resolution equi-width histogram which is based on $N = 100$ buckets. We assume that it will describe PDF(x) with enough accuracy. To build this histogram we used 10000 samples of x values that were generated using PDF(x). The histogram uses a series of obtained f_i values (series of frequencies of falling a x values in the i-th bucket) where $i = 1, \ldots, N$. It defines a probability density function:

$$f(x) = \frac{1}{h} f_i I_i(x) \wedge I_i(x) = \begin{cases} 1 \text{ if } x \text{ belongs to the } i\text{-th bucket} \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

where $h = (\max(x) - \min(x))/N$ is a width of buckets of the histogram. The probability density function $f(x)$ is presented in Fig. 1. It will be called a high resolution referential histogram.

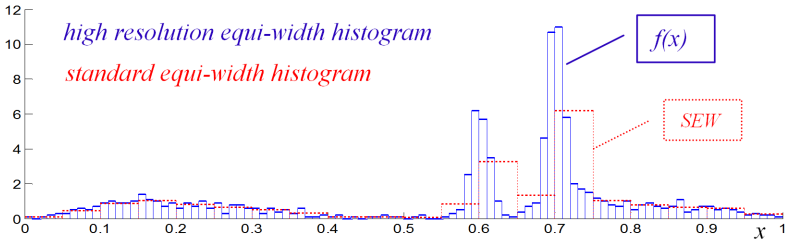

Fig. 1. Referential representation of x attribute values distribution – the high resolution equi-width histogram with $N = 100$ buckets (solid lines); *SEW* – standard equi-width histogram with $B = 20$ buckets (dotted lines) further defined in section 2.2

This histogram is named the high resolution one because other considered-below low resolution histograms will have significantly less number of buckets ($B \ll N$). The histogram presented in Fig. 1 (solid lines) will be used as temporary referential accurate distribution representation. It will be used for obtaining exact values of selectivities for any query ranges during creating standard equi-width histograms or multi-histograms.

2.2 Standard Equi-width Histogram

Let us use standard equi-width histogram (with B buckets) as a well-known typical representation of the distribution of x attribute values. It will be called *SEW*. Bucket's boundaries of *SEW* are uniformly distributed along the x domain. $B = 20$ is the assumed number of buckets in our exemplary histogram shown in Fig. 1 (dotted lines).

In the paper, we will try to find a better distribution representation (more accurate for selectivity estimations) subject to the assumed value of number of buckets (B) and taking into account an additional information about a distribution of previously processed range queries.

2.3 Exemplary of Distribution of Range Query Condition Bounds

Let us assume that we have information about a distribution of boundaries (a_j, b_j) of previously processed range queries $Q_j(a_j < x < b_j)$, where $0 \leq a_j \leq b_j \leq 1$. We assume that we have a sample – a set named $Qset$ – which consists of pairs (a_j, b_j) for $j = 1, \ldots, M$ that come from conditions of M last processed range queries. Our exemplary $Qset$ presented in Fig. 2 has $M = 20$ pairs [3]. 16 of them (circles) are highly clustered in so-called hot regions A, B, C). Zoomed parts of domain $a \times b$ were shown in Fig. 2 for presenting hot regions.

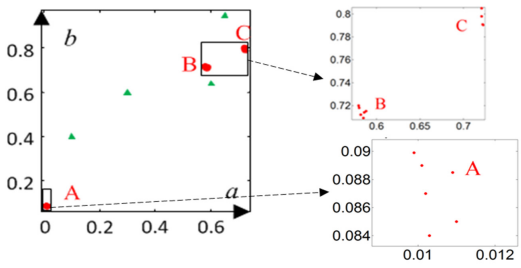

Fig. 2. $Qset$ – the set of boundaries of recently processed range query – the exemplary set of pairs (a_j, bj) for $j = 1 \ldots M(M = 20)$; A, B, C – hot regions [3]

2.4 Clustering Range Query Boundaries from Learning Set

To take into account a distribution of boundaries of query ranges we will use centers of some clusters that were built from values of a_j and b_j as some boundaries of buckets in a new histogram.

To obtain the error-optimal number of clusters we use K-fold cross validation method. In the k-th step of K-fold procedure (where $k = 1, \ldots, K$) we divide $Qset$ into a learning set $Qset\_learn_k$ and a testing one $Qset\_test_k$ [3]. $Qset\_learn_k$ will be used for obtaining some boundaries of new histogram's buckets. $Qset\_test_k$ will be used for validation of the new histogram using some selectivity estimation error metrics.

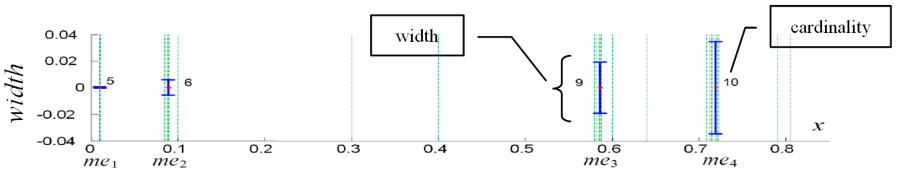

Fig. 3. Distribution of values from S_k which consists of either a or b values from $Qset\_learn_k$ – dashed lines; Medians of four accepted clusters (with their cardinality and width)

Let us assume that K equals 5 in our example. Thus $Qset\_learn_k$ consists of 16 boundaries pairs. Using $Qset\_learn_k$ we build a vector S_k. S_k has a 32 values (either a_j or b_j) that all come from $Qset\_learn_k$. Elements of S_k was presented in Fig. 3 (dashed lines).

By applying clustering procedure for S_k we may obtain a few clusters. In the example we use Fuzzy C-means algorithm (FCM) [4]. After that we eliminate some so-called weak clusters (clusters with relatively low cardinality or too wide clusters i.e. with relatively high values of standard deviation) [3]. For our example we get $C_{acc} = 4$ accepted clusters (Fig. 3).

We will use centers of accepted clusters i.e. medians $me_1, \ldots, me_{acc} = 0.0103$, 0.0888, 0.586, 0.7190 as some buckets of the new histogram. Those steps are described more detailed in [3].

2.5 Creating Equi-width-Based Multi-histogram

A new type of histogram – a multi-histogram denoted by MH – is constructed as a series of equi-width sub histograms (sH) embedded in intervals defined by the centers of clusters obtained from historical data about the distribution of boundaries of query ranges.

Due to C_{acc} clusters, we have $C_{acc} + 1$ sub-histograms. They are located between C_{acc} cluster centers, i.e. $\min(x), me_1, \ldots, me_{acc}, \max(x)$. Each equi-width sub histogram sHr is described by: s_r – a start value, e_r – an end value, B_r – a number of buckets where $r = 1, \ldots, C_{acc} + 1$ and $e_r = s_r + 1$ for $r \leq C_{acc}$, $s_1 = \min(x), e_{Cacc+1} = \max(x)$.

Let us assume that B is a given total number of bucket in the multi-histogram (i.e. in all sub-histograms). Thus:

$$B = \sum_{r=1}^{C_{acc}+1} B_r. \tag{4}$$

The multi-histogram has $B + 1$ buckets boundaries. $C_{acc} + 2$ boundaries are already defined by domains of sub-histograms, i.e. set of pairs (s_r, e_r), and $\min(x)$, and $\max(x)$. Remaining $B + 1 - (C_{acc} + 2)$ boundaries should be distributed among sub-histograms.

To obtain a final multi-histogram definition we should propose values $B_r - 1$, i.e. numbers of internal boundaries in each sub-histogram sH_r.

In this first approach to distributing locations of $B - C_{acc} - 1$ boundaries we assume that those locations should be almost uniformly distributed. Let us denote $L_r = e_r - s_r$ as a width of the sub-histogram sH_r, and $L = \max(x) - \min(x) = \sum_{r=1}^{C_{acc}+1} L_r$ as a width of the whole multi-histogram. We assume here that $B_r - 1$ should by approximately proportional to a relative width of the r-th sub-histogram:

$$B_r - 1 \approx A\frac{L_r}{L} \tag{5}$$

where A is some unknown constant.

Using (4) and (5) we may obtain A and B_r:

$$A \approx B - C_{acc} - 1, \tag{6}$$

$$B_r - 1 \approx (B - C_{acc} - 1)\frac{L_r}{L}. \tag{7}$$

In (5)–(7) we used symbol $\approx$ because B_r is a natural number so, in fact, we numerically find the optimal series of $B_1, \ldots, B_r, \ldots, B_{Cacc+1}$ and $B_r \in \mathcal{N}$ that minimizes some square evaluation function $F(B_1, \ldots, B_r, \ldots, B_{Cacc+1}) = \sum_{r=1}^{C_{acc}+1}((B - C_{acc} - 1)\frac{L_r}{L} + 1 - B_r)^2$ subject to (4).

Having B_r we may construct all sub-histograms i.e. the final multi-histogram, using values of the high-resolution referential histogram (which is shown in Fig. 1).

To evaluate an accuracy of any histogram H we use the following error metrics:

– a relative selectivity estimation error for Q (a given condition range query):

$$RelErrSel_H(a,b) = RelErrSel_H(Q(a < x < b)) = \frac{|\widehat{sel_H}(Q) - sel(Q)|}{sel(Q)} \cdot 100\%, \tag{8}$$

– a mean relative selectivity estimation error for QS (a given set of conditions):

$$MeanRelErrSel_H(QS) = \text{mean}_{(a,b)\in QS}\ \ RelErrSel_H(a,b). \tag{9}$$

$\widehat{sel_H}$ denotes an approximated selectivity value calculated with a H histogram. H is SEW (standard equi-width histogram) or MH (multi-histogram). sel denotes an exact value of selectivity calculated with the high-resolution referential histogram from Fig. 1.

Using (9) and the testing set $Qset\_test_k$ (see section 2.3) as QS we obtain: $MeanRelErrSel_{MH}(Qset\_test_k) \approx 15.8 < MeanRelErrSel_{SEW}(Qset\_test_k) \approx 31.7$. Thus, in our example the multi-histogram (MH shown in Fig. 4a) gives better accuracy (than SEW) in selectivity estimations for range query conditions from $Qset\_test_k$.

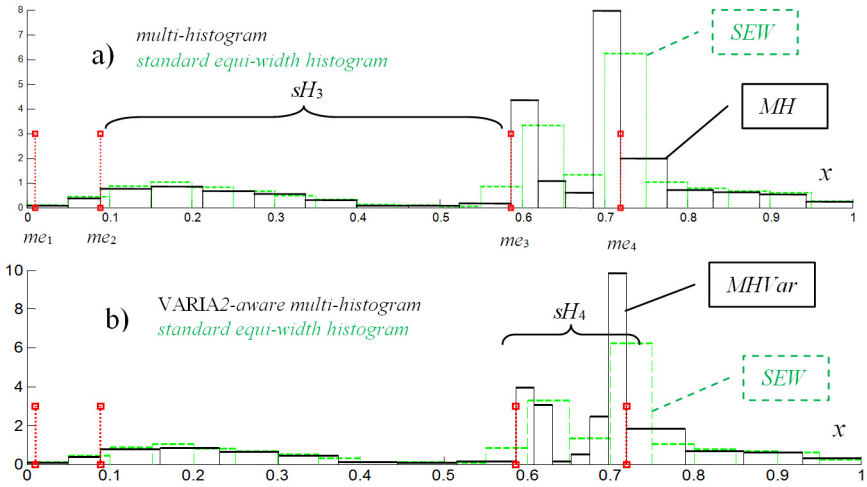

Fig. 4. *SEW* – standard equi-width histogram (dashed lines); locations of borders between sub histograms (vertical dotted lines) – series of medians; multi-histograms: a) *MH* – simple multi-histogram (solid lines), b) *MHVar* (alternatively called VARIA2-aware) – frequencies variability aware multi-histogram (solid lines) further defined in section 2.7

2.6 Improving Multi-histogram by Eliminating Boundaries from those Sub-histograms that Describe almost Uniform Distributions

To improve a multi-histogram we propose to take into account variability of frequencies (describing x distribution) in process of obtaining a distribution of numbers of internal buckets (B_r) in sub-histograms (sH_r).

In this second approach to built a multi-histogram we will assume that after finding C_{acc} we do not introduce any internal boundaries into such sub-histogram where there exist no significant changes of frequencies f_i (eq. (3) and Fig. 1) that belong to the domain of this sub-histogram. After the proposed step, such sub-histogram will have only a one bucket. So there will be more boundaries to distribute among remaining sub-histograms.

To select such sub-histogram sH_r we propose to use such condition:

$$\mathrm{VARIA1}_r = \frac{\mathrm{std}(f(x))|_{x\in[s_r,e_r]}}{\mathrm{MV}_r} \leq \mathrm{THR} \tag{10}$$

where THR is some threshold value (e.g. from $1\% \sim 10\%$), and

$$\mathrm{MV}_r = \mathrm{mean}(f(x))|_{x\in[s_r,e_r]} = \int_{s_r}^{e_r} f(x)dx, \tag{11}$$

and $\mathrm{std}(f(x))|_{x\in[s_r,e_r]} = \left(\int_{s_r}^{e_r}(f(x)-\mathrm{MV}_r)^2dx\right)^{\frac{1}{2}}$.

For the exemplary sub-histograms we obtain $(\text{VARIA1}_r)_{r=1}^{C_{acc}+1} = (0.01, 72.3, 89.3, 97.6, 114)$ what allows to select only the first sub-histogram ($r = 1$) according to the assumed threshold value (THR = 10%). This sub-histogram i.e. sH_1 will not be taken into account in further distributing of $B - C_{acc} - 1$ locations of boundaries (described in section 2.7). It will have only a one bucket.

In fact, the above-considered try of improvement of this *MH* does not change the distribution of $B - C_{acc} - 1$ locations of boundaries in our example (because the selected sub-histogram sH_1 already has only a one bucket) but it may have an impact on the distribution of boundaries locations for other distribution of x values or values in *Qset*'s elements.

2.7 Improving Multi-histogram by Increasing Resolution of Selected Sub-histograms

In this section we propose more advanced metrics of variability of distribution of frequencies f_i within a domain $[s_r, e_r]$ of some sub-histogram sH_r. Using it we may increase a resolution of some sub-histograms for high variability of frequencies.

Till now we have assumed that we take into account a distribution of query conditions i.e. a 2-dimensional distribution of pairs (a, b) (samples from this distribution are given by $Qset\_learn_k$) to find extreme boundaries of sub-histograms (series of cluster's medians). Now let us also take into account a distribution of widths or query ranges, i.e. a 1-dimensional distribution of $z = b - a$. This distribution of z will have an impact on a distribution of number of internal boundaries of sub-histograms. Such approach may allows to partially adapt a multi-histogram to a (possible in future) shifted distribution of query ranges. The exemplary discrete z values distribution (obtained from $Qset\_learn_k$) is shown in Fig. 5.

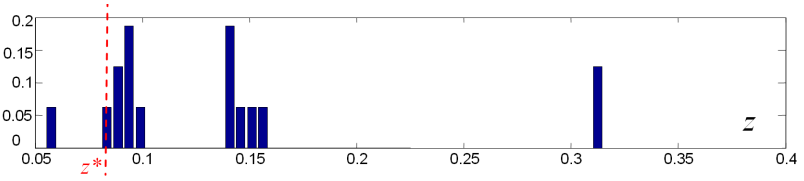

Fig. 5. The distribution of $z = b - a$, i.e. the distribution of widths of query ranges obtained from $Qset\_learn_k$

z distribution allows to obtain a window size for further analysis i.e. we chose as the window size such maximal z^* that:

$$P(z \geq z^*) = p \tag{12}$$

where p is an assumed value of confidence level.

Let us assume $p = 0.9$. Then for the considered example (i.e. the z distribution from Fig. 5) we obtain $z^* = 0.082$ (dashed line in Fig. 5).

We will use the value of z^* to determine a maximal resolution with which we will evaluate a variability of frequencies f_i (see section 2.1) within a domain of a sub-histogram.

Let us define a window function $w(t)$ with width equals z^*:

$$w(t) = \mathbf{1}(t) - \mathbf{1}(t - z^*) \tag{13}$$

where $\mathbf{1}(t)$ is the step function.

We will consider a series of queries with ranges that are included in a domain of a sub-histogram sH_r and with range lengths equal z^*. This means that we will consider window queries $Q_w(t < x < t + z^*)$ for all $t \in [s_r, e_r - z^*]$.

If we assume for a little that a sub-histogram sH_r has only a one bucket then such histogram contains only one single value equals MV_r (given by (11). Let us find a selectivity of Q_w for given t using such one-bucket sH_r:

$$sel_{1\,bck\text{-}sHr}(Q_w(t < x < t + z^*)) = \int_t^{t+z^*} \mathrm{MV}_r d\tau = \mathrm{MV}_r\ z^* = \mathrm{const}. \tag{14}$$

We may find a selectivity of Q_w for given t using $f(x)$:

$$sel(Q_w(t < x < t + z^*)) = \int_t^{t+z^*} f(\tau) w(\tau - t) d\tau \tag{15}$$

as a function convolution of f and w on interval $[t, t + z^*]$.

Let us define a scaled selectivity formula as follows:

$$\frac{sel(Q_w(t < x < t + z^*))}{z^*} = \frac{1}{z^*} \int_t^{t+z^*} f(\tau) w(\tau - t) d\tau. \tag{16}$$

We may consider the scaled $sel(Q_w)$ as a moving average filter. The result of applying the filter for the exemplary frequencies f_i (see section 2.1) and z^* equals 0.08 is shown in Fig. 6.

We introduce a new metrics of variability of frequencies within a sub-histogram sH_r (using a filter based on value of z^*) as mean selectivity estimation absolute error:

$$\begin{aligned}
&\mathrm{VARIA2}_r = \textstyle\int_{s_r}^{e_r - z^*} |sel_{1\,bck\text{-}sHr}(Q_w(t<x<t+z^*)) - sel(Q_w(t<x<t+z^*))| dt, \\
&\mathrm{VARIA2}_r = \textstyle\int_{s_r}^{e_r - z^*} |\mathrm{MV}_r\ z^* - sel(Q_w(t < x < t + z^*))| dt, \\
&\mathrm{VARIA2}_r = frac1z^* \textstyle\int_{s_r}^{e_r - z^*} |\mathrm{MV}_r - \frac{sel(Q_w(t<x<t+z^*))}{z^*}| dt.
\end{aligned} \tag{17}$$

The presented above definition of $\mathrm{VARIA2}_r$ ratio (17) is valid for $e_r - s_r \geq z^*$. For a narrow sub-histogram sH_r where $e_r - s_r < z^*$, a value of $\mathrm{VARIA2}_r$ equals 0.

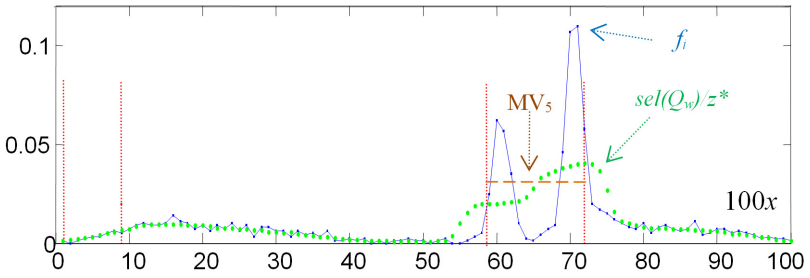

Fig. 6. Applying scaled selectivity $sel(Q_w)/z^*$ as a moving average filter ($z^* = 0.08$); frequencies f_i (connected by solid lines); averaged frequencies after applying the scaled selectivity filter (dotted lines)

Let us denote VARIA2 defined as follows:

$$\text{VARIA2} = \sum_{r=1}^{C_{acc}+1} \text{VARIA2}_r. \tag{18}$$

Using $\text{VARIA2}_r/\text{VARIA2}$ ratios we may refine the formula (5) for obtaining $B_r - 1$ as follows:

$$B_r - 1 \approx A \left(\alpha \frac{L_r}{L} + \beta \frac{\text{VARIA2}_r}{\text{VARIA2}} \right) \tag{19}$$

where α – a weight of an importance of a sub-histogram domain width and β – a weight of an importance of variability of frequencies within this sub histogram ($\alpha, \beta \geq 0$, $\alpha + \beta = 1$).

Using (4) and (19) we may obtain:

$$A \approx \frac{B - C_{acc} - 1}{\sum_{r=1}^{C_{acc}+1} \left(\alpha \frac{L_r}{L} + \beta \frac{\text{VARIA2}_r}{\text{VARIA2}} \right)}. \tag{20}$$

A multi-histogram which numbers of *SHr*'s boundaries (B_r) depend either on L_r and VARIA2_r will be called *MHVar* (or VARIA2-aware histogram).

Here, in the considered example, we propose equal impacts of the both ratios i.e. $\alpha = \beta = 1/2$. Thus using (19) and (20) and $B_r \in \mathcal{N}$ we obtain $(B_1, \ldots, B_r, \ldots, B_{Cacc} + 1) = (1, 2, 7, 6, 4)$. The new *MHVar* based on (1, 2, 7, 6, 4) is presented in Fig. 4b (solid lines). The 4-th sub-histogram of *MHVar* has 6 buckets (domain of sH_4 is a region of higher variability of frequencies f_i). It is the greater value than 4 – the number of bucket of the 4-th sub-histogram of *MH* (shown in Fig. 4a).

To evaluate *MHVar* we again use (9) and the testing set $Qset\_test_k$ (section 2.3): $MeanRelErrSel_{MHVar}(Qset\_test_k) \approx 10.1 < MeanRelErrSel_{MH}(Qset\_test_k)$. For $Qset\_test_k$ applying the VARIA2-aware histogram (*MHVar*) gives a little better selectivity estimation accuracy than applying the previously obtained multi-histogram (*MH*).

2.8 Selecting Number of Clusters of Range Boundaries to Obtain Error-Optimal Multi-histogram

K-fold cross validation method allows to find a value of the error-optimal C_{opt} amnong C values – numbers of clusters in S_k, where $C \in \mathcal{N}$ and $1 \leq C \leq B-1$ ($B-1$ because two of boundaries from all $B+1$ boundaries are already defined by $\min(x)$ and $\max(x)$). For each value of C we obtain a value of accepted clusters C_{acc} ($0 \leq Cacc \leq C$) after eliminating weak clusters.

Averaged MeanRelErrSel (which uses (9)) allows finding the optimal C_{opt} for our example as we can see in Fig. 7. We show here only values $3 \leq C \leq 9$ (not $1 \leq C \leq 19$) because for the other C values we have C_{acc} values are equal 0 (there are no accepted clusters).

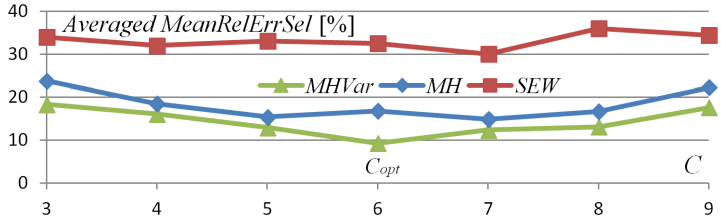

Fig. 7. K-fold cross validation results: the dependency between *Averaged MeanRelErrSel* (calculated for histograms: *MHVar, MH, SEW*) and the number of clusters equals C

For our example we obtain $C = 6$ (with corresponding $C_{acc} = 4$) as the error-optimal value, i.e. with the smallest *Averaged MeanRelErrSel* equals 10.9%.

MHVar based on $C_{opt} = 6$ clusters proved to be the most accurate representation for the considered example i.e. for the exemplary attribute distribution (given by $f(x)$ in Fig. 1) and for the exemplary distribution of boundaries of query ranges (given by $Qset$ in Fig. 2).

3 The Algorithm for Obtaining Multi-histogram

The proposed method allows obtaining the error-optimal multi-histogram *MHVar* for an arbitrary given value of B i.e. the number of *MHVar*'s buckets.

We assume that we have available $Qset$ i.e. M pairs of boundaries of last recently processed range query conditions.

The proposed method consists on the following steps:

1. Create a temporary referential representation of x attribute values distribution i.e. build a high resolution equi-width histogram which describes $f(x)$.
2. Create *SEW* – a low resolution standard equi-width histogram using the high resolution histogram.

3. For each C from $1, \ldots, B-1$ (using K-fold cross validation method) obtain an error metrics value i.e. an *Average MeanRelErrSel* which evaluates *MHVar* histogram based on C clusters:
   3.1. Cluster a learning set $Qset\_learn$ (formed from values of $Qset$) using FCM.
   3.2. Eliminate weak clusters, i.e. obtain C_{acc} accepted clusters (this determines $C_{acc}+1$ sub-histograms sH_r where $r = 1, \ldots, C_{acc}+1$).
   3.3. Assume an only one bucket in those sub-histograms sH_r where $\text{VARIA1}_r \leq \text{THR}$ (eq. (10)).
   3.4. Distribute $B - C_{acc} - 1$ locations of bucket boundaries among sub-histograms according to (19) and (20) using VARIA2_r (but omitting the sub-histograms with one bucket that were selected in 3.3).
   3.5. Having bucket's boundaries, create *MHVar* using the high resolution histogram.
   3.6. Obtain *MeanRelErrSel* for *MHVar* and *SEW* using query ranges from a testing set (i.e.: $Qset\_test = Qset \setminus Qset\_learn$).
   3.7. Aggregate values of *MeanRelErrSel*.
4. Choose this *MHVar* multi-histogram which has the smallest *Averaged MeanRelErrSel* (if *Averaged MeanRelErrSel* for *MHVar* is less than the one for *SEW*, else choose *SEW*).

The proposed method of obtaining *MHVar* is designated to be invoked during update statistics (not during on-line query processing) so it is rather not a time-critical operation (in opposite to selectivity calculation).

4 Conclusions

The proposed method of range query selectivity estimation is based on a multi-histogram – *MH*. A *MH* is such representation of an attribute values distribution which additionally takes into account a distribution of range boundaries of recently processed queries. Query workload is reflected in a division of *MH* into sub-histograms by using centers of clusters of range boundaries values. These results in creating equi-width sub-histograms embedded in *MH*. In the paper we also propose an improved multi-histogram – *MHVar* – which additionally supports an increased resolution in sub-histograms based on regions of high variability of an attribute values distribution. In *MHVar* we do not distribute buckets among the sub-histogram where attribute distribution is almost uniform. Additionally, having a knowledge of past query workload we may set a window size in some moving average filter and measure the variability of frequencies in sub-histograms. This allows refining the distribution of boundaries among selected sub-histograms embedded in *MHVar*.

In future we plan to confirm advantages of *MHVar* (in accuracy of range query selectivity estimations) against different query workload profiles in more experiments.

The future work may concentrate on improving of handling historical data about query workload. In this approach we store M last processed queries. This

may be too short-time description of a past query workload. Thus in future, we plan to use the micro clustering technique [12] which allows taking into account some impact of older queries.

Another direction of research may be considering the other type of histogram as a sub-histogram (i.e. not necessary equi-width one) like equi-high one or V-optimal one.

References

1. Augustyn, D.R.: Query-condition-aware histograms in selectivity estimation method. In: Czachórski, T., Kozielski, S., Stańczyk, U. (eds.) Man-Machine Interactions 2. AISC, vol. 103, pp. 437–446. Springer, Heidelberg (2011), 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), http://dx.doi.org/10.2478/bpasts-2014-0029
3. Augustyn, D.R.: Query selectivity estimation based on improved V-optimal histogram by introducing information about distribution of boundaries of range query conditions. In: Saeed, K., Snášel, V. (eds.) CISIM 2014. LNCS, vol. 8838, pp. 151–164. Springer, Heidelberg (2014), http://dx.doi.org/10.1007/978-3-662-45237-0_16
4. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Kluwer Academic Publishers, Norwell (1981)
5. Bruno, N., Chaudhuri, S., Gravano, L.: Stholes: A multidimensional workload-aware histogram. SIGMOD Rec. 30(2), 211–222 (2001), http://doi.acm.org/10.1145/376284.375686
6. Chen, C.M., Roussopoulos, N.: Adaptive selectivity estimation using query feedback. SIGMOD Rec. 23(2), 161–172 (1994), http://doi.acm.org/10.1145/191843.191874
7. He, Z., Lee, B.S., Wang, X.S.: Proactive and reactive multi-dimensional histogram maintenance for selectivity estimation. J. Syst. Softw. 81(3), 414–430 (2008), http://dx.doi.org/10.1016/j.jss.2007.03.088
8. Ioannidis, Y.: The history of histograms (abridged). In: Proc. of VLDB Conference (2003)
9. Khachatryan, A., Müller, E., Stier, C., Böhm, K.: Sensitivity of self-tuning histograms: Query order affecting accuracy and robustness. In: Ailamaki, A., Bowers, S. (eds.) SSDBM 2012. LNCS, vol. 7338, pp. 334–342. Springer, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-31235-9_22
10. Luo, J., Zhou, X., Zhang, Y., Shen, H.T., Li, J.: Selectivity estimation by batch-query based histogram and parametric method. In: Proceedings of the Eighteenth Conference on Australasian Database, ADC 2007, vol. 63, pp. 93–102. Australian Computer Society, Inc., Darlinghurst (2007), http://dl.acm.org/citation.cfm?id=1273730.1273741
11. Srivastava, U., Haas, P.J., Markl, V., Kutsch, M., Tran, T.M.: Isomer: Consistent histogram construction using query feedback. In: Proceedings of the 22nd International Conference on Data Engineering, ICDE 2006, pp. 39–51. IEEE Computer Society, Washington, DC (2006), http://dx.doi.org/10.1109/ICDE.2006.84
12. Zhang, T., Ramakrishnan, R., Livny, M.: Birch: An efficient data clustering method for very large databases. SIGMOD Rec. 25(2), 103–114 (1996), http://doi.acm.org/10.1145/235968.233324