STUDIA INFORMATICA 2002
Volume 23 Number 4 (51)

Dariusz Rafał AUGUSTYN
Politechnika Śląska, Instytut Informatyki

METODA ESTYMACJI JĄDROWEJ W SZACOWANIU SELEKTYWNOŚCI ZAPYTAŃ

Streszczenie. W artykule dokonany został przegląd wybranych metod estymacji nieparametrycznej, wykorzystanych do szacowania współczynnika selektywności zapytań. Artykuł koncentruje się głównie na metodzie estymacji jądrowej, użytej do przybliżania nieznanej funkcji gęstości, opisującej rozkład wartości atrybutu tablicy bazy danych. Estymowana funkcja gęstości pozwala na oszacowanie selektywności zapytań wykorzystywanej przez optymalizator zapytań. Pokazana jest koncepcja wykorzystania metody estymacji jądrowej dla wyznaczania selektywności łącznie dla zbioru atrybutów, bez zakładania niezależności tychże, na podstawie wielowymiarowego estymatora jądrowego.

METHOD OF KERNEL ESTIMATION IN APPROXIMATION OF QUERY SELECTIVITY

Summary. The article presents a survey of methods of nonparametric estimation used for estimation of query selectivity. The article mainly focuses on kernel estimation used for approximation of unknown density function of distribution of values from database table attribute. An approximation of density function lets calculate a query selectivity, used by database query optimizer. The paper presents multidimensional kernel estimator used for calculation of common query selectivity for set of attribute without the assumption of attributes independence.

1. Wstęp

Jednym z zadań programu serwera bazy danych jest efektywna realizacja zapytań. Wymaga ona wcześniejszego opracowania przez serwer bazy danych optymalnej metody realizacji pod względem czasu otrzymania odpowiedzi. Fazą realizacji zapytania,

poprzedzającą pobranie danych jest wytworzenie tzw. planu wykonania, czyli wyboru istniejących struktur danych, pozwalającego na możliwie najszybsze otrzymanie odpowiedzi, będącej wynikowym zbiorem wierszy. Zadanie stworzenia planu wykonania jest realizowane przez podsystem systemu zarządzania bazami danych (SZBD) zwany optymalizatorem zapytań. Wybór najlepszego planu wykonania między innymi opiera się na przybliżonym określeniu rozmiaru zbioru wierszy spełniających kryteria zapytania, pochodzących z tablicy lub tablic, których to zapytanie dotyczy. Określenie przybliżonej liczby wierszy, spełniających kryteria zapytania dla każdej tablicy występującej w zapytaniu, pozwala na wybór najlepszej metody dostępu do danych (np. sekwencyjny przegląd stron bazy danych z wierszami danej tablicy albo dostęp do odpowiednich stron z wykorzystaniem drzewa indeksowego), mierzonej jak najmniejszą liczbą odczytów z bazy danych (liczbą pobrań stron bazy danych z pamięci masowych).

Selektywność $Sel(Q)$ prostego zapytania Q definiuje się jako:

$$Sel(Q) = \frac{\text{Liczba wierszy spełniających warunek zapytania } Q}{\text{Liczba wszystkich wierszy tablicy}} \tag{1}$$

W przedstawionej definicji zakłada się, że proste zapytanie Q dotyczy pojedynczej tablicy bazy danych. Selektywność jest wartością z przedziału $\langle 0, 1 \rangle$ i można ją interpretować jako prawdopodobieństwo wyboru wiersza spełniającego kryteria zapytania Q z całego zbioru wszystkich wierszy danej tablicy.

Jeżeli zapytanie dotyczy kilku tablic bazy danych rozkładane ono jest przez optymalizator SZBD na podzapytania proste. Dotyczą one pojedynczych tablic bazy danych lub tablic tymczasowych, będących tablicami pośrednimi powstającymi w ramach planowanych etapów realizacji zapytania złożonego.

Selektywność dowolnego zapytania nie jest na ogół dokładnie znana przed wykonaniem zapytania. Jest ona oszacowywana na podstawie pewnych, zgromadzonych wcześniej danych statystycznych, opisujących cechy wartości w kolumnach tablic. W dużych bazach, dane statystyczne, ze względu na czasochłonność operacji ich tworzenia, nie są wyznaczane z wykorzystaniem wszystkich wierszy danej tablicy, ale w oparciu o próbę losową [4].

Dane statystyczne opisujące rozkład wartości w kolumnie mogą mieć charakter parametryczny (np. wartość średnia, odchylenie standardowe, stosunek unikalnych wartości do liczby wierszy) lub nieparametryczny (np. histogram). Opis parametryczny wymaga bardzo mało zasobów w postaci przestrzeni (mała zajętość pamięci) w porównaniu do rozważanego w artykule opisu nieparametrycznego, ale wykorzystanie estymacji parametrycznej nieznanego rozkładu wartości w kolumnie jest na ogół obarczone większym błędem.

W artykule zaprezentowano jedną z metod estymacji nieparametrycznej, tj. estymację jądrową, zastosowaną do szacowania współczynnika selektywności. Pokazano pewne zalety estymacji jądrowej w stosunku do estymacji histogramowej, powszechnie stosowanej w SZBD.

2. Związek selektywności zapytania i funkcji gęstości rozkładu prawdopodobieństwa wartości atrybutu relacji

W relacyjnym modelu danych tablica bazy danych, nazywana relacją R, zawiera wiersze zwane krotkami r (relacja jest zbiorem krotek, tzn. $R = \{r\}$). Kolumny tablic nazywane są atrybutami A_i. Zbiór atrybutów określa schemat relacji $R\ [A_1, A_2, \ldots A_n]$.

W artykule rozważane są atrybuty o ciągłej dziedzinie. Po pewnych modyfikacjach można zastosować wyniki poniższych rozważań także do atrybutów o dziedzinie przeliczalnej (dyskretnej, z określoną relacją porządku).

Rozkład wartości atrybutu A_i o ciągłej dziedzinie można opisać wzorem:

$$A_i: F(x) = P(A_i < x) = \int_{-\infty}^{x} f(x)dx, \tag{2}$$

gdzie f jest funkcją gęstości prawdopodobieństwa, a F dystrybuantą rozkładu zmiennej losowej A_i.

W artykule rozważane są zapytania proste Q (A_i, a, b), kierowane do pojedynczej relacji R, które zawierają warunek logiczny w postaci kryterium zakresu dla wartości atrybutu A_i. Wynikiem realizacji takiego zapytania Q jest zbiór wierszy $\{r\}$ taki, że $a \le r.A_i \le b$. Uwzględniając wzory (1) i (2) można znaleźć zależność pomiędzy funkcją gęstości prawdopodobieństwa A_i i selektywnością zapytania Q:

$$Sel(Q) = \int_{a}^{b} f(x)dx\ . \tag{3}$$

Na mocy wzoru (3) selektywność można wyznaczyć na podstawie funkcji gęstości $f(x)$. Ze względu na brak dokładnej znajomości f można wykorzystać przybliżenie $\hat{f}(x)$, czyli estymator f z próby losowej $X_1, X_2, \ldots, X_n$.

3. Ocena błędu estymacji funkcji gęstości prawdopodobieństwa

Miarą jakości estymacji może być scałkowany błąd średniokwadratowy *ISE* (ang. integrated square error):

$$ISE\left(\hat{f}\right) = \int_{-\infty}^{+\infty} (\hat{f}(x) - f(x))^2 dx. \tag{4}$$

Błąd *ISE* wyznaczany jest na podstawie konkretnej próby losowej. Poprzez uśrednienie błędu *ISE* po wszystkich możliwych próbach (wartość oczekiwana *ISE*) można wprowadzić uniwersalną miarę jakości estymacji, tzn. średni scałkowany błąd średniokwadratowy *MISE* (ang. mean integrated square error) :

$$MISE\left(\hat{f}\right) = E\left(\int_{-\infty}^{+\infty} (\hat{f}(x) - f(x))^2 dx\right), \tag{5}$$

gdzie $E(Z)$ oznacza wartość średnią zmiennej losowej Z.

Praktyczne wyznaczenie błędu *MISE* jest trudne. Wartość błędu zależy nie tylko od nieznanej funkcji f, ale i od liczności próby n. Jednak dla dużej liczności próby, gdy znajdują zastosowanie centralne twierdzenia graniczne, można znaleźć asymptotyczne przybliżenie *MISE*, tzn. asymptotycznie aproksymowany średni scałkowany błąd średniokwadratowy *AMISE* (ang. asymptotically approximated mean integrated square error).

4. Estymacja histogramowa

W SZBD często stosowaną metodą estymacji nieparametrycznej jest estymacja histogramowa, zdefiniowana następująco:

$$\hat{f}(x) = \frac{n_i}{n} I_{(c_i, c_{i+1})}(x), \tag{6}$$

gdzie:

$\hat{f}(x)$ – estymator f z próby $X_1, X_2, \ldots, X_{n_i}$,

n – liczność próby,

c_i – punkty podziału dziedziny X wyznaczające granice przedziałów histogramu,

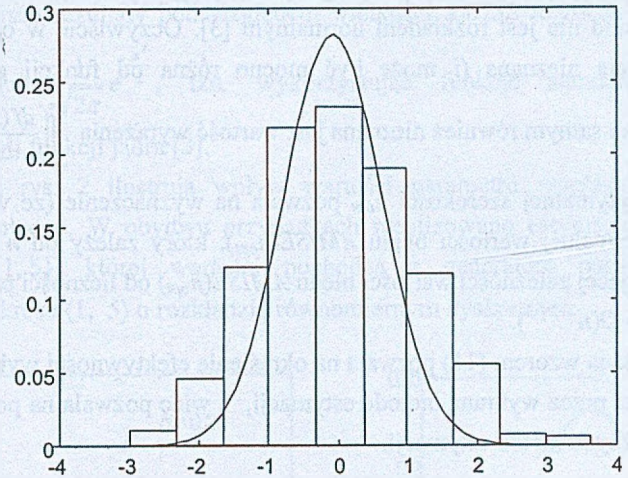
n_i – liczba elementów próby o wartościach z przedziału (c_i, c_{i+1}),

I – funkcja przynależności x do zbioru S, zdefiniowana jako $I(x)_S = \begin{cases} 1 \text{ dla } x \in S \\ 0 \text{ dla } x \notin S \end{cases}$.

W wielu zastosowaniach (w tym również w niektórych optymalizatorach SZBD) wykorzystywane są histogramy o jednakowej szerokości h wszystkich przedziałów (ang. equi-width histograms), tzn. takie histogramy, że dla każdego i spełnione jest $c_{i+1} - c_i = h$.

Na rys. 1 można porównać wykres funkcji gęstości standardowego rozkładu normalnego $N(0, 1)$ z wykresem histogramu o jednakowej szerokości przedziału równej $2/3$, histogramu

estymującego funkcję gęstości standardowego rozkładu normalnego, wykonanego na podstawie próby tysiąca elementów, pochodzących z generatora liczb pseudolosowych o rozkładzie normalnym.



Rys. 1. Estymacja histogramowa standardowego jednostkowego rozkładu normalnego
Fig. 1. Histogram estimation of standard normal distribution

Błąd *AMISE* estymacji funkcji *f* wykorzystującej histogram o stałej szerokości przedziału *h*, sporządzonego dla próby o odpowiednio dużej liczności *n*, wynosi:

$$AMISE(h) = \frac{1}{nh} + \frac{h^2}{12}\int_{-\infty}^{+\infty}\left(\frac{df(x)}{dx}\right)^2 dx. \tag{7}$$

Wzór (7) dotyczy różniczkowalnych funkcji *f*, dla których spełnione jest $\int_{-\infty}^{+\infty}\left(\frac{df(x)}{dx}\right)^2 dx < \infty$.

Dla ustalonej liczności próby i funkcji *f* można (bez znajomości *f*) znaleźć optymalną wartość *h*, tzn. takie h_{opt}, dla którego *AMISE* jest najmniejsze, tzn:

$$\frac{d}{dh} AMISE(h_{opt}) = 0. \tag{8}$$

Stąd (wzór (7) i (8)) optymalna szerokość przedziału histogramu wynosi:

$$h_{opt} = \left(\frac{6}{n\int_{-\infty}^{+\infty}\left(\frac{df(x)}{dx}\right)^2 dx}\right)^{\frac{1}{3}}. \tag{9}$$

Dla przykładu, na podstawie wzoru (9), dla rozkładu normalnego o standardowym odchyleniu estymowanym przez s, optymalna szerokość wynosi:

$$h_{opt\_normal} \approx 3{,}486\, s\, n^{-1/3}. \tag{10}$$

Taka technika znajdywania h optymalnego (wzór (10)) daje dobre rezultaty, nawet gdy estymowany rozkład nie jest rozkładem normalnym [3]. Oczywiście w ogólnym wypadku funkcja f pozostaje nieznana (i może być mocno różna od funkcji gęstości rozkładu normalnego), a tym samym również nieznana jest wartość wyrażenia $\int_{-\infty}^{+\infty} \left(\frac{df(x)}{dx} \right)^2 dx$.

Znalezienie optymalnej szerokości h_{opt} pozwala na wyznaczenie (ze wzorów (7) i (9)) odpowiedniej optymalnej wartości błędu $AMISE(h_{opt})$, który zależy od n i f. Można więc znaleźć rząd malejącej zależności wartości błędu $AMISE(h_{opt})$ od liczności próby n:

$$AMISE(h_{opt}) = O(n^{-2/3}). \tag{11}$$

Zależność podana wzorem (11) pozwala na określenie efektywności wykorzystania próby o zadanej liczności przez wybraną metodę estymacji, a więc pozwala na porównywanie pod tym względem różnych metod estymacji.

5. Estymator jądrowy

Estymator jądrowy (ang. kernel estimation) $\hat{f}$ funkcji gęstości f z próby $X_1, X_2, \ldots, X_n$, określony jest jako:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=0}^{n} K\left(\frac{x - X_i}{h} \right), \tag{12}$$

gdzie:

h – parametr wygładzenia nazywany też szerokością okna lub szerokością pasma,

K – funkcja nazywana jądrem (pewna funkcja gęstości prawdopodobieństwa).

Estymacja jądrowa polega na superpozycji funkcji jądra w punktach próby. Parametr h decyduje o „spłaszczeniu" funkcji jądrowych, wchodzących do sumy tworzącej wyrażenie opisujące estymator.

Funkcja jądra $K(x)$ posiada następujące własności:

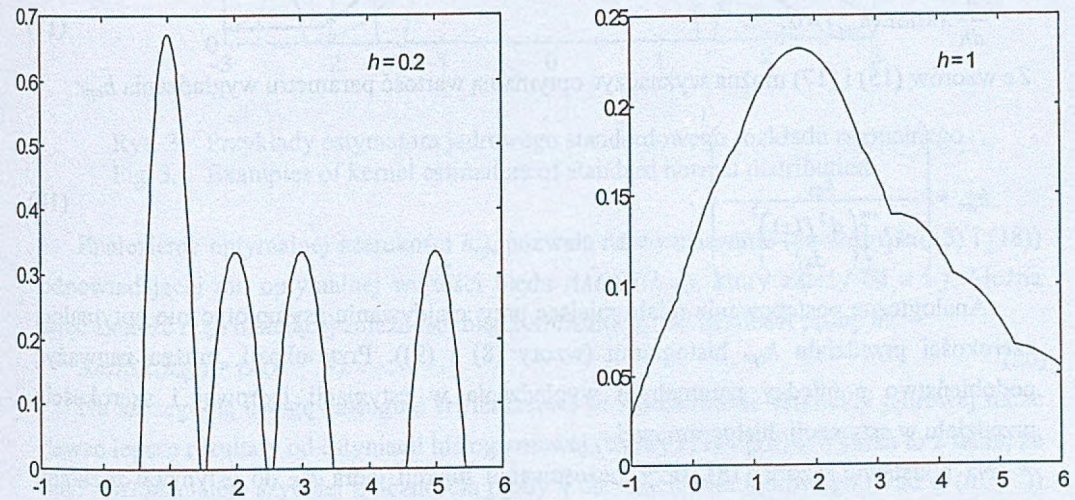
$$\int_{-\infty}^{+\infty} K(x)dx = 1, \quad K(x) \geq 0, \quad \int_{-\infty}^{+\infty} xK(x)dx = 0. \tag{13}$$

Optymalną pod względem wielkości błędu estymacji $AMISE$ (pozwalającą na uzyskanie możliwie najmniejszych wartości $AMISE$) funkcją jądra jest tzw. jądro Epanechnikova [1, 3]:

$$K_E(x) = \begin{cases} \frac{3}{4}(1-x^2) \text{ dla } |x| \le 1, \\ 0 \text{ w przeciwnym razie.} \end{cases} \quad (14)$$

Niewiele gorsze rezultaty pod względem wielkości *AMISE* uzyskuje się stosując jądro gaussowskie $K_G(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$, tzn. wykorzystując funkcję standardowego rozkładu normalnego w roli funkcji jądra [3].

Przykłady z rys. 2 ilustrują wpływ wartości parametru wygładzania h na postać estymatora jądrowego. W obydwu przypadkach zrealizowano estymację z jądrem K_E dla próby $\{1, 2, 3, 1, 5\}$, której wartości pochodzą z generatora pseudolosowego liczb naturalnych z zakresu $\langle 1, 5 \rangle$ o rozkładzie równomiernym dyskretnym.



Rys. 2. Przykłady estymacji jądrowej dla parametru wygładzania h równego 0.2 i 1
Fig. 2. Examples of kernel estimation for values 0.2 and 1 of smooth parameter h

Asymptotycznie aproksymowany średni scałkowany błąd średniokwadratowy *AMISE* dla estymacji jądrowej wynosi:

$$AMISE(h) = \frac{h^4}{4} k_2 \int_{-\infty}^{+\infty} \left(\frac{d^2 f(x)}{dx^2} \right)^2 dx + \frac{1}{nh} k_{02}, \quad (15)$$

gdzie k_2 i k_{02} są własnościami zastosowanej funkcji jądra:

$$k_{02} = \int_{-\infty}^{+\infty} K(x)^2 dx, \; k_2 = \int_{-\infty}^{+\infty} x^2 K(x) dx \neq 0. \quad (16)$$

Wzór (15) dotyczy funkcji f dwukrotnie różniczkowalnych, dla których spełnione jest $\int_{-\infty}^{+\infty}\left(\frac{d^2 f(x)}{dx^2}\right)^2 dx < \infty$.

Jądro Epanechikova K_E (wzór (14)) jest jądrem asymptotycznie optymalnym, tzn. daje minimalny błąd *AMISE* dany wzorem (15) w klasie funkcji spełniających założenia dane wzorami (13) i (16) (przy założeniu o dużej liczności próby). Znalezienie K_E sprowadza się do minimalizacji po funkcjach K funkcjonału *AMISE* (K). Jądro Epanechnikova $K = K_E$ spełnia warunek o skończonej wartości momentu drugiego rzędu, dając $k_{02} = 0.6$ i $k_2 = 0.2$ z zależności (16).

Niezależnie od zastosowanego jądra (niekoniecznie K_E), z wzoru (15) można znaleźć asymptotycznie optymalną wartość parametru wygładzania h_{opt}, zależną od liczności próby n (i pewnych cech jądra) oraz nieznanej funkcji gęstości f, rozwiązując:

$$\frac{d}{dh} AMISE(h_{opt}) = 0. \tag{17}$$

Ze wzorów (15) i (17) można wyznaczyć optymalną wartość parametru wygładzania h_{opt}:

$$h_{opt} = \left(\frac{k_{02}}{n k_2^2 \int_{-\infty}^{+\infty}\left(\frac{d^2 f(x)}{dx^2}\right)^2 dx} \right)^{\frac{1}{5}}. \tag{18}$$

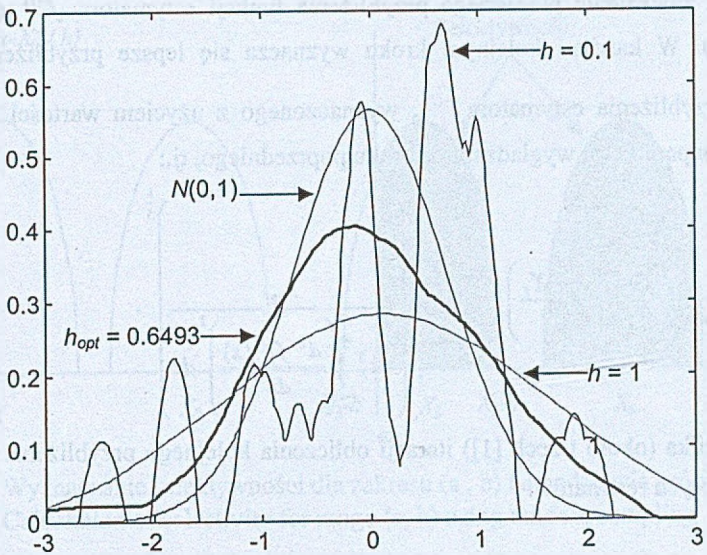
Analogiczne postępowanie miało miejsce przy znajdywaniu asymptotycznie optymalnej szerokości przedziału h_{opt} histogramu (wzory (8) i (9)). Przy okazji, można zauważyć podobieństwo pomiędzy parametrem wygładzania w estymacji jądrowej i szerokością przedziału w estymacji histogramowej.

Na podstawie wzoru (18), przy zastosowaniu funkcji jądra K_E do estymacji rozkładu normalnego o dowolnym odchyleniu standardowym, optymalna wartość parametru wygładzania wynosi:

$$h_{opt\_normal} \approx 1.05\ s\ n^{-1/5}, \tag{19}$$

gdzie s jest estymatorem odchylenia standardowego liczonym z próby.

Rysunek 3 przedstawia wykresy estymatorów jądrowych standardowego rozkładu normalnego dla parametrów wygładzania h: 0.1 , 0.6493, 1. Jako funkcję jądra zastosowano funkcję Epanechnikova. Estymatory jądrowe zostały zbudowane na podstawie 30-elementowej próby losowej. Wartość optymalnego wygładzania $h_{opt} \approx 0.6493$ została uzyskana ze wzoru (19).



Rys. 3. Przykłady estymatora jądrowego standardowego rozkładu normalnego
Fig. 3. Examples of kernel estimators of standard normal distribution

Znalezienie optymalnej szerokości h_{opt} pozwala na wyznaczenie (ze wzorów (15) i (18)) odpowiadającej mu optymalnej wartości błędu $AMISE(h_{opt})$, który zależy od n i f. Można więc znaleźć rząd malejącej zależności błędu $AMISE(h_{opt})$ od liczności próby n:

$$AMISE(h_{opt}) = O(n^{-4/5}). \tag{20}$$

Na szczególną uwagę zasługuje stwierdzenie, że zastosowanie estymacji jądrowej może dawać lepsze rezultaty od estymacji histogramowej (wzory (11) i (20)). Wynika to z faktu, że błąd *AMISE* maleje szybciej z licznością próby n dla estymatora jądrowego (rząd $O(n^{-4/5})$) niż dla estymatora histogramowego (rząd $O(n^{-2/3})$), więc wykorzystanie próby w metodzie estymacji jądrowej jest lepsze. Dlatego być może pojawią się implementacje algorytmów estymacji jądrowej w nowych wersjach eksperymentalnych SZBD.

Wyznaczenie parametru wygładzania h, dla którego błąd *AMISE* byłby najmniejszy (wzór (18)), wymaga znajomości nieznanej funkcji f. W praktyce dobre rezultaty osiąga się zakładając na potrzeby wyznaczania h_{opt}, że f może być opisana gęstością rozkładu normalnego (reguła sprowadzania do rozkładu normalnego - ang. normal scale rule) nawet, jeśli tak nie jest.

Jeżeli funkcja f byłaby jednak znacząco różna od rozkładu normalnego, można zastosować pewną technikę iteracyjną (ang. direct plug-in rule) [1]. Wyznaczenie $h_{opt}^{(1)}$ z wykorzystaniem założenia o rozkładzie normalnym jest pierwszym krokiem tego algorytmu

i pozwala na znalezienie pierwszego przybliżenia funkcji estymatora $\hat{f}^{(1)}$ ($h = h_{opt}^{(1)}$ we wzorze (12)). W każdym kolejnym kroku wyznacza się lepsze przybliżenie $h_{opt}^{(i+1)}$ na podstawie przybliżenia estymatora $\hat{f}^{(i)}$, wyznaczonego z użyciem wartości $h_{opt}^{(i)}$, będącej przybliżeniem parametru wygładzania z kroku poprzedniego, tj.:

$$\hat{f}^{(i)}(x) = \frac{1}{nh_{opt}^{(i)}} \sum_{i=0}^{n} K\left(\frac{x - X_i}{h_{opt}^{(i)}}\right), \quad h_{opt}^{(i+1)} = \left(\frac{k_{02}}{nk_2^2 \int_{-\infty}^{+\infty} \left(\frac{d^2 \hat{f}^{(i)}(x)}{dx^2} \right)^2 dx} \right)^{\frac{1}{5}}. \tag{21}$$

W praktyce kilka (około trzech [1]) iteracji obliczenia kolejnego przybliżenia wystarczy do uzyskania dobrych rezultatów.

5.1. Metody efektywnego obliczania selektywności w metodzie estymacji jądrowej

Wykorzystanie estymacji jądrowej wiąże się m.in. z koniecznością rozwiązania problemów takich jak:

- przechowywanie elementów próby oraz ich pielęgnacji (aktualizacja próby po aktualizacji bazy danych),
- efektywne przeglądanie elementów licznej próby podczas estymacji selektywności dla konkretnego zapytania.

Wyrażenie na selektywność można zapisać w postaci:

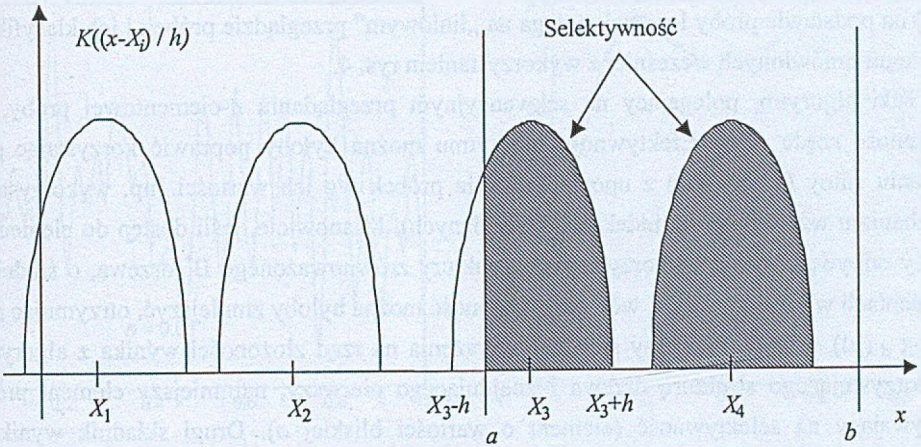
$$Sel(Q) = \int_a^b \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) dx = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \int_a^b K\left(\frac{x - X_i}{h}\right) dx. \tag{22}$$

Po podstawieniu $t = (x - X_i) / h$ do wzoru (22) selektywność wyraża się następująco:

$$Sel(Q) = \frac{1}{n} \sum_{i=1}^{n} \int_{(a-X_i)/h}^{(b-X_i)/h} K(t)dt. \tag{23}$$

Jeśli wykorzystane zostanie jądro Epanechnikova ($K = K_E$), przez co uwzględniona będzie pewna cecha funkcji jądra K_E (tzn. fakt, że funkcja ta jest dodatnio określona tylko w pewnym przedziale, a poza nim jest równa zero), to często nie wszystkie elementy próby będą wykorzystane w algorytmie wyznaczającym selektywność konkretnych zapytań.

Dla niektórych próbek X_i odpowiadające im wartości całki oznaczonej z $K(t)$ we wzorze (23) będą wynosić 0 lub 1.



Rys. 4. Wyznaczanie selektywności dla zakresu ⟨a , b⟩ na podstawie próby losowej
Fig. 4. Calculation of selectivity for range ⟨a, b⟩ using random sampling

Rysunek 4 pokazuje w jaki sposób jądro oparte na elemencie próby X_i uwzględniane jest w wyznaczaniu selektywności zapytania, którego kryterium dotyczy przynależności do przedziału o krańcach a i b $(a \le b)$.

Pewne próbki, jak np. X_1 i X_2, nie wpływają na selektywność, tzn. wnoszą 0 w sumie ze wzoru (23). Ogólnie są nimi próbki X_i, dla których spełniony jest warunek:

$$(X_i + h < a) \vee (X_i - h > b) \,. \tag{24}$$

Dla próbki X_4 z rys. 4 część dziedziny funkcji jądra opartej na tej próbce, dla której funkcja ta jest niezerowa, całkowicie zawiera się w przedziale ⟨a, b⟩. Takie właśnie próbki, dla których spełniony jest warunek:

$$(X_i - h > a) \wedge (X_i + h < b), \tag{25}$$

wnoszą 1 do sumy ze wzoru (23).

Obliczenia całek wchodzących do sumy ze wzoru (23) wymagają próbki nie spełniające żadnego z warunków (24) i (25), jak np. X_3 na rys. 4. Ze względu na efektywność obliczeń całkowanie funkcji gęstości w przedziale można sprowadzić do wyznaczenia różnicy wartości dystrybuanty w górnej i dolnej granicy przedziału. Dla jądra Epanachnikova dystrybuanta wynosi:

$$F_E(x) = \begin{cases} \frac{1}{4}(1 - x^3) \text{ dla } |x| \le 1, \\ 0 \text{ w przeciwnym razie.} \end{cases} \tag{26}$$

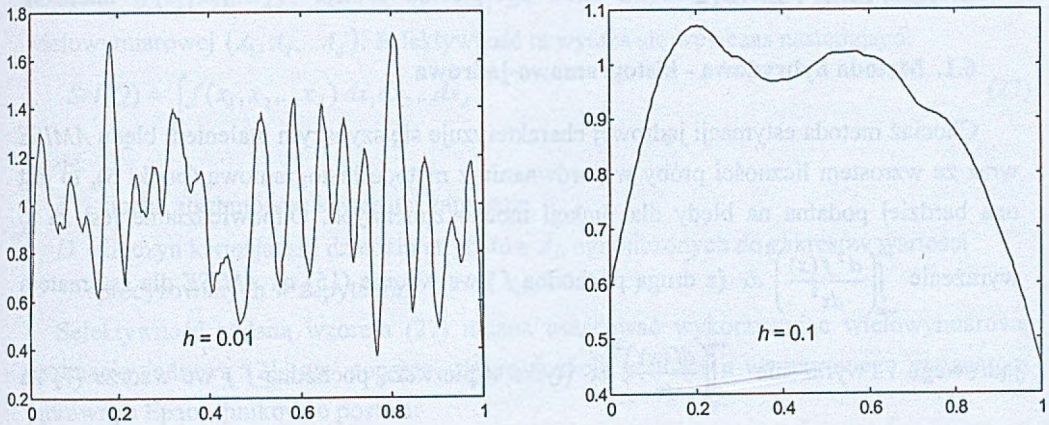
W najprostszej wersji algorytm wyznaczania selektywności, czyli liczenia sumy ze wzoru (23) na podstawie próby losowej, polega na „liniowym" przeglądzie próbek i ich klasyfikacji wg reguł omówionych wcześniej z wykorzystaniem rys. 4.

Taki algorytm, polegający na sekwencyjnym przeglądaniu n-elementowej próby, ma złożoność rzędu $O(n)$. Efektywność algorytmu można byłoby poprawić korzystając przy liczeniu sumy (wzór (23)) z uporządkowania próbek wg ich wartości (np. wykorzystując mechanizm wzorowany na indeksach bazy danych). Mianowicie, jeśli dostęp do elementów próby odbywałby się z wykorzystaniem struktury zrównoważonego B^+-drzewa, o średnio d elementach w każdym węźle, wówczas złożoność można byłoby zmniejszyć, otrzymując rząd $O(\log_d(n)) + O(k)$. Pierwszy składnik wyrażenia na rząd złożoności wynika z algorytmu wykorzystującego strukturę drzewa i znajdującego pierwszy, najmniejszy element próby, wpływający na selektywność (element o wartości bliskiej a). Drugi składnik wynika z algorytmu "liniowego" przeglądania listy kolejnych, uporządkowanych rosnąco elementów próby (liści drzewa), aż do „największego" elementu, jeszcze wpływającego na wartość selektywności (elementu o wartości bliskiej b). Zmienna k ($k \leq n$) jest liczbą elementów próby uwzględnianych przy wyznaczaniu selektywności dla zadanych a i b.

5.2. Problem brzegowy

Błąd przybliżenia f dla estymatora jądrowego jest duży w punktach, w których następują duże zmiany wartości funkcji gęstości f (tzn. decyduje o tym duża wartość wyrażenia $\int_{-\infty}^{+\infty} \left(\frac{d^2 f(x)}{dx^2} \right)^2 dx$ we wzorze (15) na *AMISE* w tych punktach).

Sytuacja taka występuje w szczególności dla rozkładów, o których wiadomo, że funkcja gęstości jest dodatnio określona w tylko w pewnym przedziale i zerowa poza nim lub w sytuacjach, gdzie następuje skokowa (lub prawie skokowa) zmiana wartości funkcji z/do wartości 0 na krańcach przedziału. Takie cechy f powodują powstawanie niekorzystnie małych wartości estymatora (w stosunku do faktycznych wartości dokładnej funkcji gęstości f) w okolicach krańców przedziału. Efekt ten nazywany jest problemem brzegowym (ang. boundary problem). Jego ilustrację może stanowić rys. 5, na którym zaprezentowano estymator jądrowy przybliżający funkcję gęstości rozkładu jednostajnego na odcinku $\langle 0, 1)$, na podstawie 300-elementowej próby losowej. Problem błędów estymacji na krańcach przedziału bardziej ujawnia się dla większych wartości parametru wygładzania (np. na rys. 5 - duże odstępstwo od wartości 1 na krańcach przedziału przy $h = 0.1$).



Rys. 5. Przykłady estymacji jądrowej funkcji gęstości rozkładu równomiernego na odcinku ⟨0, 1) dla parametru wygładzania h równego 0.01 i 0.1

Fig. 5. Examples of kernel estimation of density function of equi-distribution ⟨0, 1) for values 0.01 and 0.1 of smooth parameter h

Zbyt małe wartości estymatora jądrowego rozkładu równomiernego na krańcach przedziału wynikają z braku próbek pochodzących z lewego sąsiedztwa dla kresu dolnego i z prawego dla kresu górnego. Wraz ze wzrostem h wpływ obecności sąsiadujących próbek na wartość estymatora w danym punkcie jest większy. Stąd też jednostronny brak sąsiedztwa na krańcach uwidacznia się w postaci malejących wartości estymatora wraz ze wzrostem h.

Metodą eliminacji tego efektu może być utworzenie zbioru dodatkowych "sztucznych" próbek $(\{X_{li}\} \cup \{X_{pi}\})$ poprzez symetryczne odbicie względem krańca przedziału części próbek położnych na prawo dla lewego krańca l (utworzenie $\{X_{li}\}$) i części próbek położonych na lewo dla prawego krańca p (utworzenie $\{X_{pi}\}$), tj:

$$\underset{X_i \le l + D_l}{\forall} X_{li} := l - (X_i - l), \quad \underset{X_i \ge p - D_p}{\forall} X_{pi} := p - (X_i - p),$$

gdzie D_l i D_p wyznaczają granice zbiorów próbek do "powielania" dla lewego i prawego krańca. Uzyskany w ten sposób skorygowany estymator, ważny tylko w przedziale ⟨l, p⟩ (chociaż niektóre próbki go tworzące są spoza tego przedziału), może nie spełniać warunku

$$\int_l^p \hat{f}(x)dx = 1.$$

Inną metodę eliminacji problemu brzegowego można scharakteryzować jako pewną modyfikację funkcji jądra, której postać nie jest jednakowa dla wszystkich próbek (wartości funkcji jądra są większe dla próbek w okolicach krańców przedziału; funkcja jądra nie spełnia wszystkich cech funkcji gęstości).

6. Kierunki rozwoju

6.1. Metoda hybrydowa - histogramowo-jądrowa

Chociaż metoda estymacji jądrowej charakteryzuje się szybszym maleniem błędu *AMISE* wraz ze wzrostem liczności próby w porównaniu z metodą histogramową (punkt 5), to jest ona bardziej podatna na błędy dla funkcji mocno zmiennych. Odpowiedzialne jest za to wyrażenie $\int_{-\infty}^{+\infty}\left(\frac{d^2 f(x)}{dx^2}\right)^2 dx$ (z drugą pochodną f) we wzorze (15) na *AMISE* dla estymatora jądrowego i wyrażenie $\int_{-\infty}^{+\infty}\left(\frac{df(x)}{dx}\right)^2 dx$ (tylko z pierwszą pochodną f) we wzorze (7) na *AMISE* dla estymatora histogramowego. Stąd estymator jądrowy jest bardziej wrażliwy na zmienność f.

Ponadto (jak już wspomniano w punkcie 5.1), w punktach dużych zmian wartości funkcji gęstości również z powodu wyrażenia $\int_{-\infty}^{+\infty}\left(\frac{d^2 f(x)}{dx^2}\right)^2 dx$, będącego miarą tzw. szorstkości funkcji f, estymacja jądrowa charakteryzuje się dużą bezwzględną wartością błędu *AMISE*.

Na podstawie tych dwóch przesłanek powstała idea połączenia metod estymacji histogramowej i jądrowej [1]. Polega ona na wykryciu punktów dużych zmian funkcji gęstości (oszacowaniu na podstawie próby). W ten sposób można wyodrębnić przedziały (pewnego histogramu o przedziałach różnej szerokości), w których nie następują duże zmiany funkcji gęstości. W ramach każdego takiego przedziału można później zastosować estymator jądrowy, po uprzednim wyznaczeniu optymalnej wartości parametru wygładzania (indywidualnie dla każdego z przedziałów).

6.2. Zastosowanie estymatora jądrowego wielowymiarowego

Estymacja jądrowa może znaleźć zastosowanie do wyznaczania selektywności zapytań, w których warunek selekcji dotyczy kilku atrybutów relacji. Wyznaczając selektywność optymalizatory SZBD wyznaczają selektywność dla każdego atrybutu, a następnie mnożą je przez siebie (tak jak mnoży się prawdopodobieństwa zdarzeń niezależnych dla uzyskania prawdopodobieństwa koniunkcji tych zdarzeń). Domyślnie zakłada się wzajemną niezależność zmiennych losowych odpowiadających atrybutom występującym w kryterium zapytania.

Założenie o niezależności często nie jest spełnione. Aby więc uzyskać lepszą dokładność w szacowaniu selektywności, należałoby uzyskać funkcję gęstości wielowymiarowego

rozkładu $f(x_1, x_2, ...x_d)$, łącznie opisującego zbiór atrybutów, czyli rozkładu zmiennej wielowymiarowej $(A_1, A_2, ...A_d)$. Selektywność ta wyraża się wówczas następująco:

$$Sel(Q) = \int_D f(x_1, x_2, ...x_d)\, dx_1 dx_2 ... dx_d , \tag{27}$$

gdzie:

d – liczba atrybutów w kryterium zapytania,

D – iloczyn kartezjański dziedzin atrybutów A_i, ograniczonych do zakresów wartości sprecyzowanych w zapytaniu.

Selektywność zadaną wzorem (27) można oszacować wykorzystując wielowymiarową estymację jądrową [2], np. poprzez użycie funkcji gęstości d-wymiarowego estymatora jądrowego Epanechnikova o postaci:

$$K_E(x_1, ..., x_d) = \begin{cases} \left(\frac{3}{4}\right)^d \frac{1}{h_1 ... h_d} \prod_{i=1}^{d} \left(1 - \left(\frac{x_i}{h_i}\right)^2\right) & \text{dla } \left|\frac{x_i}{h_i}\right| < 1 \\ 0 & \text{w przeciwnym wypadku} \end{cases}, \tag{28}$$

gdzie h_i jest parametrem wygładzania wyznaczonym dla i-tego wymiaru.

Wartość parametru wygładzania wyraża się wzorem:

$$h_i = \sqrt{5} s_i n^{\frac{1}{-d+4}}, \tag{29}$$

gdzie:

s_i – odchylenie standardowe wartości atrybutu A_i, liczone z próby,

n – rozmiar próby.

7. Zakończenie

Artykuł ma charakter przeglądowy. Celem artykułu jest pokazanie kierunku rozwoju SZBD, jakim jest zastosowanie metod statystyki matematycznej w optymalizacji zapytań przy wyznaczaniu selektywności zapytań. Metody nieparametryczne są dość powszechnie stosowane przez optymalizatory SZBD głównie poprzez użycie metod histogramowych (histogramy equi-width, equi-high). Jednak SZBD nie wspomagają administratorów w zakresie koniecznych decyzji dotyczących wartości istotnych parametrów estymacji, jak np. liczba przedziałów histogramu [4] (a przecież asymptotycznie optymalne wartości niektórych parametrów są wyznaczalne analitycznie).

W pracy zaproponowano zastosowanie estymacji jądrowej w optymalizacji zapytań, pokazując pewne zalety takiego podejścia w stosunku do idei wykorzystania estymacji

histogramowej. Estymatory jądrowe zgodnie z aktualną wiedzą autora prawdopodobnie nie znalazły jeszcze zastosowań w komercyjnych optymalizatorach SZBD.

Systemy komercyjne na ogół wyznaczają selektywność pewnych zapytań, w których warunek logiczny dotyczy kilku kolumn danej tablicy, poprzez mnożenie selektywności wyznaczanych odrębnie dla każdej z kolumn. Jest tu domyślnie zastosowane założenie o niezależności zmiennych losowych opisujących rozkłady wartości w kolumnach. Założenie to bardzo często bywa jednak nieprawdziwe. Zastosowanie wielowymiarowego estymatora jądrowego do estymacji wielowymiarowej funkcji gęstości dałoby z pewnością lepsze oszacowanie selektywności w zapytaniach tego typu.

Artykuł stanowi materiał do dalszych badań nad zastosowaniem estymatorów jądrowych w szacowaniu wartości selektywności zapytań, wykorzystywanych do wypracowywania optymalnych planów wykonania zapytań. Kierunki dalszych prac związane będą z minimalizacją zajętości pamięci, wykorzystywanej na przechowywanie próbek. Planowane jest zbadanie metod aproksymacji estymatora jądrowego funkcji gęstości (lub dystrybuanty). Wówczas możliwe będzie przechowywanie mniejszej ilości danych, tzn. jedynie danych o węzłach aproksymacji, zamiast danych o wartościach wszystkich próbek. Badania dotyczyć będą efektywności wyznaczania selektywności w oparciu o estymator jądrowy z wykorzystaniem skompresowanej próby.

LITERATURA

1. Blohsfeld B., Korus D., Seeger B.: A Comparison of Selectivity Estimator for Range Queries on Metric Attributes. ACM SIGMOD'99, 1999.
2. Gunopulos D., Kollios G., Tsotras V.J.: Approximating Multi-Dimensional Aggregate Range Queries Over Real Attributes. ACM SIGMOD 2000, Dallas 2000.
3. Gajek L., Kałuszka M.: Wnioskowanie statystyczne. WNT, Warszawa 1996.
4. Leverenz L.: ORACLE 8 Concepts. ORACLE Corp. 1998.

Recenzent: Dr hab. inż. Stanisław Wołek, Prof. Pol. Rzeszowskiej

Wpłynęło do Redakcji 8 lutego 2002 r.

Abstract

The article presents some methods of statistics in a query optimization, which are or could be used by DBMS optimizers. The paper is a short survey of methods non-parametric estimation of an unknown density function of distribution of values of a database table attribute. Histogram and kernel estimators was described and compared. Estimators were used to approximate a value of query selectivity. The paper shows formulas, which present how an error of estimation depends on a sample size. Formulas for an optimal parameter for histogram and kernel estimators were shown, too. Advantages and disadvantages of kernel estimators are presented. The paper presents a multidimensional kernel estimator used for calculation of query selectivity for set of attributes without an often assumption of attributes independence on which bases most of commercial DBMS.