STUDIA INFORMATICA 2013
Volume 34 Number 2A (111)

Dariusz R. AUGUSTYN
Politechnika Śląska, Instytut Informatyki

ZASTOSOWANIE PREDYKCJI ROZKŁADU WARTOŚCI ATRYBUTU W CELU POPRAWY DOKŁADNOŚCI ESTYMACJI SELEKTYWNOŚCI ZAPYTAŃ

Streszczenie. Parametr selektywności jest wykorzystywany w procesie optymalizacji zapytań. Uzyskanie selektywności wymaga nieparametrycznego estymatora rozkładu wartości atrybutu, tj. histogramu. Histogramy są tworzone w ramach procesu aktualizacji statystyk. Dla dużych baz danych aktualizacja statystyk jest wykonywana raczej rzadko, np. tylko w momentach małego obciążenia systemu. To powoduje, że histogramy nie opisują aktualnego rozkładu danych. Aby uzyskać bardziej aktualne histogramy, powinno się zastosować mechanizm predykcji rozkładu. Pozwoli to na bardziej dokładną estymację selektywności. W niniejszym artykule zaproponowano metodę ekstrapolacji rozkładu wartości atrybutów. Metoda ta dokonuje predykcji momentów szukanego, ekstrapolowanego rozkładu. W celu jego wyznaczenia opisywana metoda wykorzystuje zasadę maksimum entropii z uwzględnieniem wartości momentów znalezionych w ramach procedury predykcji.

Słowa kluczowe: estymacja selektywności zapytań, histogram, zasada maksimum entropii rozkładu, ewolucja funkcji gęstości prawdopodobieństwa, predykcja szeregów czasowych

APPLYING PREDICTION OF ATTRIBUTE VALUE DISTRIBUTION FOR IMPROVEMENT OF QUERY SELECTIVITY ESTIMATION ACCURACY

Summary. A selectivity parameter is needed in query optimization process. Obtaining the query selectivity requires a non-parametric estimator of attribute value distribution, i.e. a histogram. Histograms are produced during update statistics process. For large databases the update statistics process is performed rather seldom, e.g. only during time of low workload of a system. This results that histograms do not describe actual data distribution. To obtain a more accurate histogram, a prediction mechanism should be introduced. This results obtaining a more accurate estimation of selectivity.

The method of extrapolation of attribute value distribution is proposed in this paper. This method predicts moments of the extrapolated distribution. It uses the maximum entropy principle for obtaining the extrapolated distribution subject to the predicted values of the distribution moments.

Keywords: query selectivity estimation, histogram, maximum entropy principle, evolution of probability density function, time series prediction

1. Wprowadzanie

Wykonanie zapytania przez System Zarządzania Bazą Danych (SZBD) jest poprzedzone jego analizą, którą przeprowadza tzw. optymalizator zapytań. W tej fazie przetwarzania, zwanej fazą przygotowania (ang. *prepare phase*), następuje wypracowanie sposobu realizacji zapytania (ang. *execution plan*). Spośród wielu potencjalnych metod realizacji wybierana jest metoda optymalna pod względem szacowanego kosztu realizacji. Koszt ten jest głównie mierzony liczbą pobrań danych z pamięci masowej, gdzie zdeponowane są dane (liczba pobrań z dysku jednostek alokacji pamięci). Oszacowanie kosztu jest poprzedzone przybliżonym określeniem ilości danych, które spełniają kryteria zapytania (tzn. spełniają warunek selekcji zapytania). Służy temu parametr zwany selektywnością (ang. *query selectivity*). Selektywność dla zapytań jednotablicowych to stosunek liczby wierszy spełniających kryteria zapytania do całkowitej liczby wierszy. Selektywność można również określić jako prawdopodobieństwo wylosowania wiersza spełniającego kryterium zapytania w losowaniu bez zwracania wierszy z tablicy. Dla zapytań zakresowych Q (ang. *range query*), w których warunek selekcji $a \leq X \leq b$ jest określony na atrybucie X z ciągłą dziedziną wartości, selektywność wyraża się wzorem:

$$sel(Q(a \leq X \leq b)) = \int_a^b f(x)dx\,, \tag{1}$$

gdzie $f(x)$ to funkcja gęstości prawdopodobieństwa rozkładu wartości X.

Z powyższego wynika, że na potrzeby oszacowania selektywności wymagane jest użycie nieparametrycznego estymatora funkcji gęstości, opisującego rozkład wartości atrybutu. Najczęściej w takiej roli w SZBD stosowane są histogramy, przykładowo – histogram *equi-width* o stałej szerokości podprzedziałów.

Histogramy są tworzone/aktualizowane w ramach procesu tzw. aktualizacji statystyk. Dla dużych baz danych jest to proces czasochłonny i oczywiście nie jest wykonywany na bieżąco (tzn. z każdą zmianą danych). Statystyki są aktualizowane na ogół w momentach zmniejszonej aktywności eksploatowanego systemu informatycznego, tzn. w chwilach mniejszego

„operacyjnego" obciążenia SZBD. Częstokroć są to chwile, których wystąpienia charakteryzują się regularnością (np. weekendy, pory nocne).

Optymalizator w procesie analizy zapytania korzysta ze statystyk (w tym histogramów) utworzonych ostatnio. Oczywiście zmiana danych na ogół pociąga za sobą zmianę rozkładu wartości, stąd histogramy powoli z upływem czasu tracą swoja aktualność (w stosunku do danych źródłowych, które opisują), a selektywność wyznaczona z ich użyciem staje się coraz bardziej niedokładna.

Można więc zadać pytanie: czy byłaby możliwa ekstrapolacja postaci histogramu (estymacja postaci funkcji gęstości rozkładu w niedalekiej przyszłości), gdyby koszty obsługi takiego programowego mechanizmu ekstrapolacji były mniejsze niż realizacja ponownej aktualizacji statystyki? Istotnym składnikiem kosztu są rozmiary metadanych potrzebnych do realizacji mechanizmu predykcji. Powinny być to oczywiście rozmiary niewielkie. Aby mechanizm stanowił alternatywę w stosunku do klasycznej aktualizacji statystyki na podstawie bazy danych, procedura predykcji powinna raczej wykorzystywać potencjalnie wolne moce obliczeniowe, a nie opierać się na przetwarzaniu danych z bazy danych. Reasumując, klasyczna aktualizacja statystyk to głównie utylizacja pamięci masowej (dyski), a predykcja postaci rozkładu to głównie utylizacja CPU.

W niniejszym artykule zaproponowano metodę predykcji postaci rozkładu. Metoda zakłada następujące wstępne etapy służące do strojenia parametrów metody:

- etap określenia minimalnego, wystarczająco dokładnego opisu rozkładu za pomocą momentów rozkładu rzędu $1…K$ (rozdziały 2, 3),
- etap określenia modelu predykcji wartości momentu r-tego rzędu (dla $r = 1…K$) w przyszłości, tj. dla chwili o indeksie $t + 1$, przy znajomości wartości momentów w chwilach poprzednich: t, $t – 1$,… (rozdział 4).

Chwile o indeksach t, $t – 1$,… oznaczają równoodległe momenty czasowe, w których nastąpiły kolejne aktualizacje statystyk. Chwila o indeksie $t + 1$ określa moment w czasie najbliższej planowanej aktualizacji statystyk w przyszłości.

Sama metoda estymacji postaci rozkładu w przyszłości na dowolną chwilę $\tau \in (\tau_t, \tau_{t+1})$ polega na realizacji następujących etapów:

- użycie uzyskanych poprzednio modeli predykcji do znalezienia nowych wartości momentów rozkładu na chwilę o indeksie $t + 1$ (rozdział 4),
- interpolacja wartości momentów rozkładu w chwili τ (rozdział 5),
- wykorzystanie wartości momentów w chwili τ do otrzymania estymatora rozkładu w chwili τ, czyli **uzyskanie szukanego histogramu na chwilę τ w przyszłości** (rozdział 2).

Ostatni z wymienionych etapów wykorzystuje zasadę maksimum entropii informacyjnej, tj. zakłada wyznaczenie postaci rozkładu o największej entropii przy ograniczeniach nałożonych na wartości momentów szukanego rozkładu.

W rozpatrywanej metodzie rozpatruje się takie rozkłady, dla których istnieją momenty rozkładu rzędu 1…*K*.

2. Estymacja funkcji rozkładu prawdopodobieństwa z wykorzystaniem znanych wartości momentów rozkładu oraz zasady maksimum entropii

W ramach niniejszego rozdziału zostanie przedstawiona metoda określająca sposób odtworzenia funkcji gęstości rozkładu (a dokładniej wybranego nieparametrycznego estymatora funkcji gęstości prawdopodobieństwa – histogramu *equi-width*) na podstawie znajomości wartości momentów (kilku początkowych rzędów) oraz zastosowania zasady maksimum entropii.

Problem można sformułować jako „znalezienie wartości ciągu (p_i), odpowiadających częstości wystąpień zmiennej X w podprzedziałach (o stałej szerokości w), których środki określone są przez zadane wartości rosnącego ciągu (x_i), gdzie $i = 1\ldots N$, a N oznacza liczbę podprzedziałów histogramu. Zakładając, że rozłączne podprzedziały histogramu pokrywają całą dziedzinę wartości X, można sformułować następującą równość:

$$\sum_{i=1}^{N} p_i = 1. \tag{2}$$

Załóżmy, że znane są momenty rozkładu $m^{(r)}$ rzędu r, gdzie $r = 1\ldots K$. Pozwala to na sformułowanie następującego układu K równań liniowych:

$$\begin{aligned} &\sum_{i=1}^{N} x_i p_i = m^{(1)} \\ &\quad\ldots \\ &\sum_{i=1}^{N} x_i^r p_i = m^{(r)} \\ &\quad\ldots \\ &\sum_{i=1}^{K} x_i^K p_i = m^{(K)}. \end{aligned} \tag{3}$$

Warunki (2) i (3) można syntetycznie sformułować następująco:

$$\boldsymbol{A}\boldsymbol{P}^T = \boldsymbol{M}, \tag{4}$$

gdzie:

$$\boldsymbol{A}=\begin{bmatrix} 1 & 1 & ... & 1 \\ x_1 & x_2 & ... & x_N \\ ... & ... & ... & ... \\ x_1^K & x_2^K & ... & x_N^K \end{bmatrix}, \boldsymbol{P}=[p_1 \ \cdots \ p_N], \ \boldsymbol{M}=\begin{bmatrix} 1 \\ m^{(1)} \\ ... \\ m^{(K)} \end{bmatrix}. \tag{5}$$

Dodatkowo jako obowiązujące można przyjąć następujące nierówności, wynikające z ogólnych własności prawdopodobieństwa:

$$\underset{i=1...N}{\forall} 0 \le p_i \le 1. \tag{6}$$

Entropia S rozkładu dyskretnego $\{(x_i, p_i)\}$ jest określona następująco:

$$S(p_1,...,p_N) = -\sum_{i=1}^{N} p_i \ln(p_i). \tag{7}$$

Zasada maksimum entropii rozkładu [1, 2, 3] orzeka, że przy przyjętych ograniczeniach, np. wynikających z określonych wartości momentów rozkładu, najbardziej prawdopodobny jest taki rozkład $\hat{p}_i$, dla którego wartość entropii jest największa, tj.:

$$(\hat{p}_1,...,\hat{p}_N) = \arg\left\{ \sup_{p_1,...,p_N} (S(p_1,...,p_N)) \right\}. \tag{8}$$

Uwzględniając powyższe informacje, zadanie odtworzenia rozkładu można sprowadzić do zadania optymalizacji, tj. znalezienia minimum N-argumentowej funkcji F (minus entropii) z ograniczeniami wyrażonymi formułami (4) i (6). Argumentami funkcji $F = – S$ są p_i dla $i = 1... N$.

3. Przykłady estymacji funkcji rozkładu prawdopodobieństwa z wykorzystaniem zasady maksimum entropii

Celem eksperymentów omówionych poniżej będzie oszacowanie wartości maksymalnego rzędu momentów (wartość K we wzorach (3) i (5)) niezbędnych do wystarczającej dokładności estymacji danego rozkładu.

W ramach oceny dokładności estymacji funkcji rozkładu wg omówionej powyżej metody zostaną przedstawione dwa przykłady estymacji funkcji gęstości.

3.1. Przykład 1 – rozkład jednomodalny

Przykład 1 pokazuje estymację empirycznego rozkładu zmiennej, której wartości zostały uzyskane z generatora liczb pseudolosowych o rozkładzie Gaussa $N(2, 0{,}75)$.

Rysunek 1 przedstawia histogram źródłowy sporządzony na podstawie 10 000-elementowej próby losowej. Jest to histogram o stałej szerokości podprzedziałów $w = 0{,}15$, z liczbą $N = 50$ podprzedziałów, obejmujący dziedzinę wartości X z przedziału [-2, 5,5].

Rys. 1. Źródłowy histogram opisujący rozkład oryginalny
Fig. 1. Source histogram – the original distribution

Dla oceny dokładności estymacji zaproponowano prostą metrykę błędu – wskaźnik nazywany dalej ErrChi2based – tzn. średni kwadrat względnego odchylenia prawdopodobieństw rozkładu wynikowego $\hat{p}_i$ i rozkładu źródłowego p_i (nazwany błędem rekonstrukcji rozkładu), tj.:

$$\text{ErrChi2based} = \frac{1}{N}\sum_{j=1}^{N} e_j , \tag{9}$$

gdzie składnik e_j jest określony następująco:

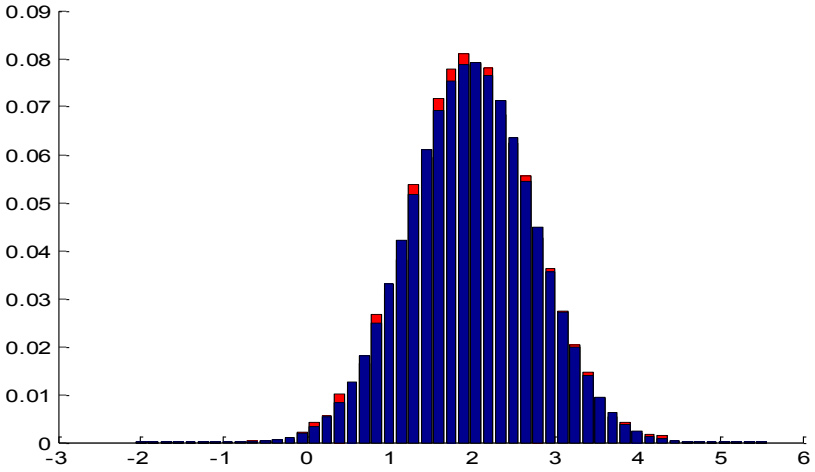
$$e_j = \begin{cases} \dfrac{(\hat{p}_j - p_j)^2}{p_j} & \text{dla} \quad p_j \neq 0 \\ \dfrac{(\hat{p}_j - p_j)^2}{\hat{p}_j} & \text{dla} \quad p_j = 0 \wedge \hat{p}_j \neq 0 \\ 0 & \text{dla} \quad p_j = 0 \wedge \hat{p}_j = 0. \end{cases} \tag{10}$$

Przyjmijmy następujące kryterium akceptacyjne estymacji:

$$\text{ErrChi2based} \leq 0{,}001 . \tag{11}$$

Wyniki eksperymentów zrealizowanych z użyciem programu Matlab dla przykładu 1 pokazują, że nawet dla bardzo małych wartości K estymacja spełnia założone kryterium dokładności, określone formułą (11). Dla $K = 2$ ($m^{(1)} = 1{,}996915$ i $m^{(2)} = 4{,}55533$) wartość ErrChi2based wyniosła zaledwie $8{,}27 \cdot 10^{-5}$. Rysunek 2 pokazuje histogram wynikowy (kolor niebieski) uzyskany przez minimalizację funkcji F (maksymalizację entropii), przy zadanych

wartościach momentów rzędu 1. i 2. Fragmenty histogramu źródłowego (kolor czerwony) również zostały przedstawione.



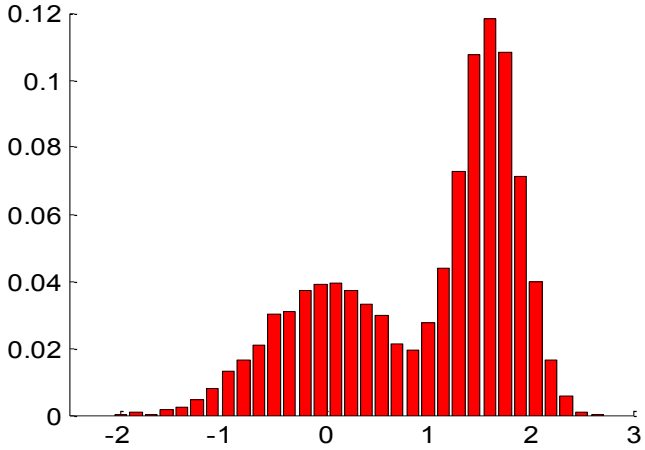
Rys. 2. Histogram wynikowy (kolor niebieski) – rozkład odtworzony na podstawie znajomości momentów rozkładu rzędu 1. i 2.
Fig. 2. Resulting histogram (blue color) – the distribution reconstructed on the basis of the 1st distribution moment and the 2nd one.

3.2. Przykład 2 – rozkład dwumodalny

Przykład 2 pokazuje estymację rozkładu empirycznego zmiennej, której wartości zostały uzyskane z generatora liczb pseudolosowych określonego następującą funkcją gęstości prawdopodobieństwa (superpozycja dwóch rozkładów Gaussa):

$$f(x) = 4/10 \cdot \mathrm{PDF}(N(0, 0{,}6)) + 6/10 \cdot (\mathrm{PDF}(N(1{,}6, 0{,}3)), \tag{12}$$

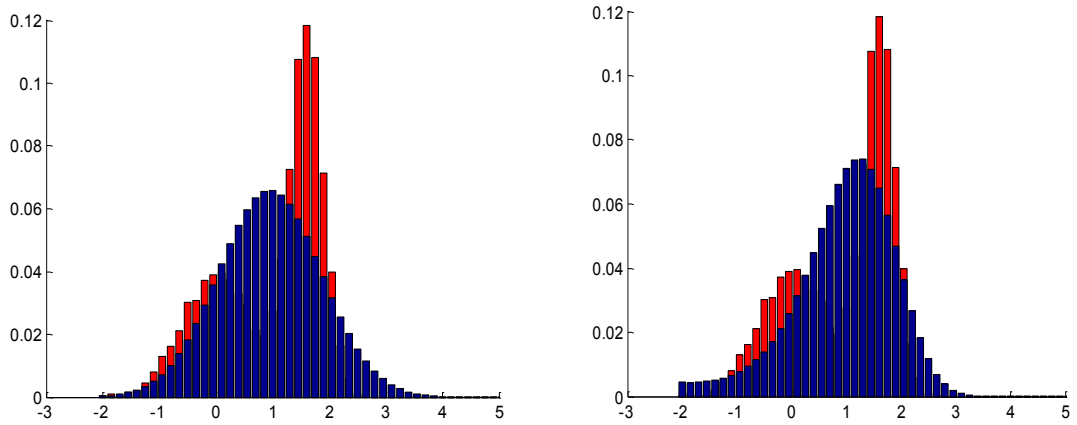
gdzie PDF($N(m, \sigma)$) oznacza funkcję gęstości prawdopodobieństwa rozkładu normalnego.



Rys. 3. Źródłowy histogram opisujący rozkład oryginalny będący superpozycją 2 klastrów Gaussa – wzór 12
Fig. 3. Source histogram – the original distribution based on 2 Gaussian clusters given by formula 12

Rysunek 3 przedstawia histogram sporządzony na podstawie 10 000-elementowej próby losowej. Jest to histogram o stałej szerokości podprzedziałów równej $w = 0{,}15$, z liczbą $N =$

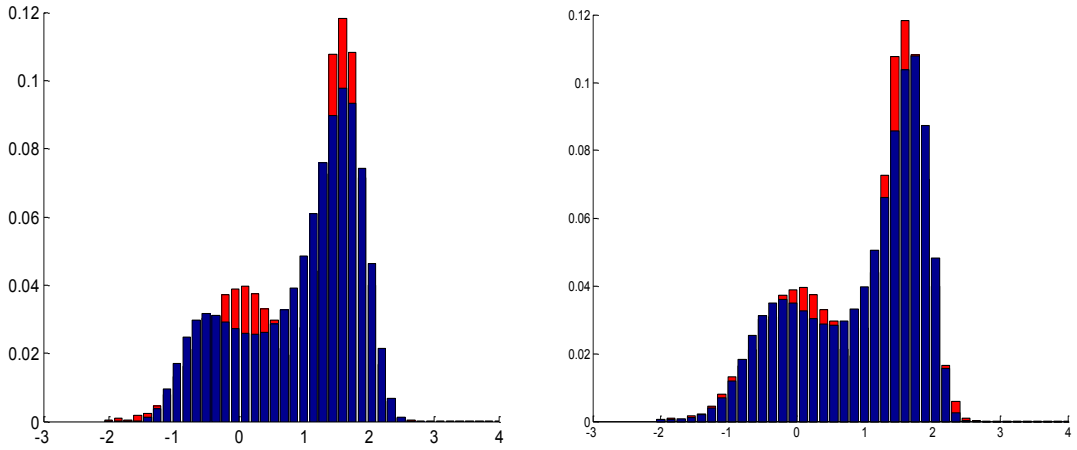
50 podprzedziałów, obejmujący dziedzinę wartości X z przedziału [-2, 5,5]. Na rys. 3-7 dziedzina X została zawężona dla zwiększenia przejrzystości rysunków, ponieważ wartości histogramu są poza nią równe zeru.



a) b)

Rys. 4. Wynikowy histogram (kolor niebieski) – przybliżenie sporządzone na podstawie: a) momentów rozkładu rzędu 1. i 2. (K = 2; ErrChi2based = 0,022167), b) momentów rozkładu rzędów od 1. do 3. (K = 3; ErrChi2based = 0,016472)

Fig. 4. Resulting histogram (blue color) – the estimation based on: a) the 1st and 2nd moments of distribution (K = 2; ErrChi2based = 0.022167), b) the 1st ÷ 3rd moments of distribution (K = 3; ErrChi2based = 0.016472)



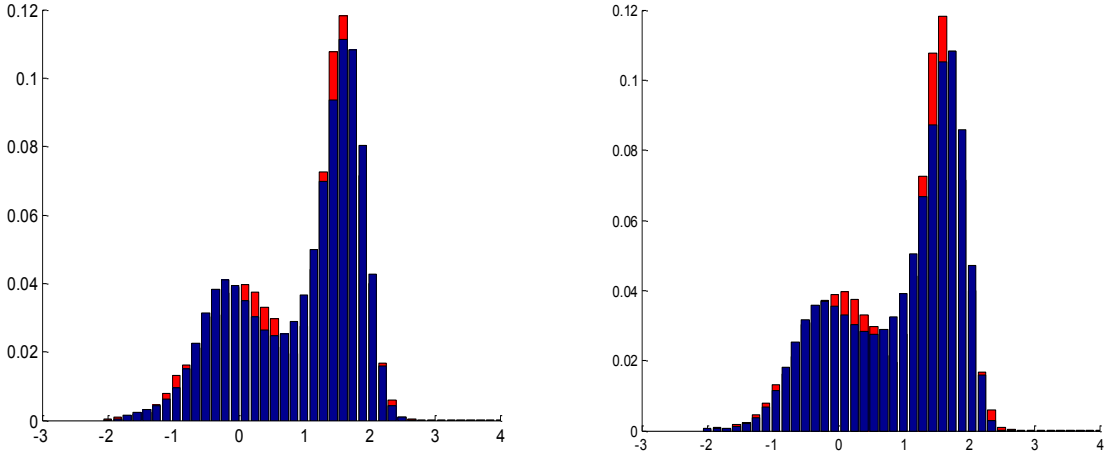
a) b)

Rys. 5. Wynikowy histogram (kolor niebieski) – przybliżenie sporządzone na podstawie: a) momentów rozkładu rzędów od 1. do 4. (K = 4; ErrChi2based = 0,00172), b) momentów rozkładu rzędów od 1. do 5. (K = 5; ErrChi2based = 8,0892e-004)

Fig. 5. Resulting histogram (blue color) – the estimation based on: a) the 1st ÷ 4th moments of distribution (K = 4; ErrChi2based = 0.00172), b) the 1st ÷ 5th moments of distribution (K = 5; ErrChi2based = 8.0892e-004)

Rysunki 4-7 prezentują wynikowy histogram (kolor niebieski) na tle histogramu źródłowego (kolor czerwony) dla różnych wartości K = 2, 3, 4, 5, 6, 7, 8, 12, czyli dla różnej uwzględnionej liczby momentów rozkładu źródłowego. Oczywiście wśród wymienionych najgorsze przybliżenie wystąpi dla K = 2 (rys. 4a), a najlepsze dla K = 12 (rys. 7b). Uzyskane

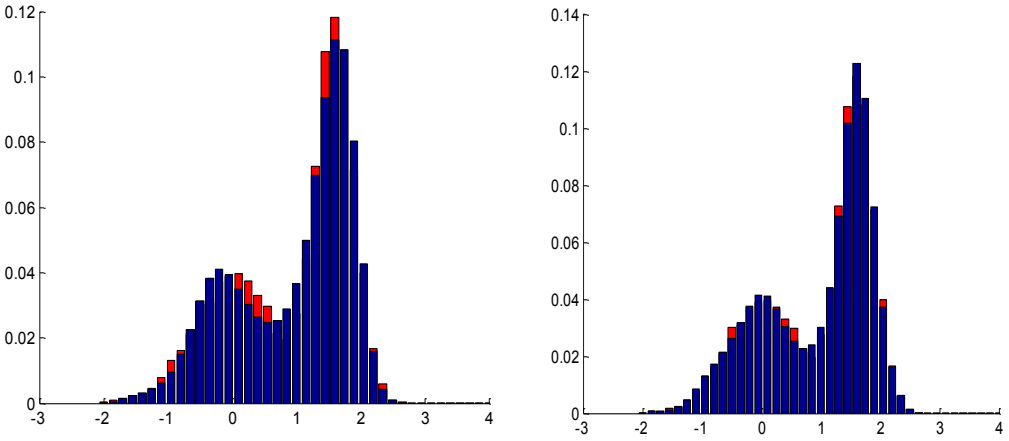
wartości błędów rekonstrukcji ErrChi2based rozkładu w zależności od *K* przedstawiono na rys. 8.



a) b)

Rys. 6. Wynikowy histogram (kolor niebieski) – przybliżenie sporządzone na podstawie: a) momentów rozkładu rzędów od 1. do 6. (*K* = 6; ErrChi2based = 8,01289e-004), b) momentów rozkładu rzędów od 1. do 7. (*K* = 7; ErrChi2based = 7,25156e-004)

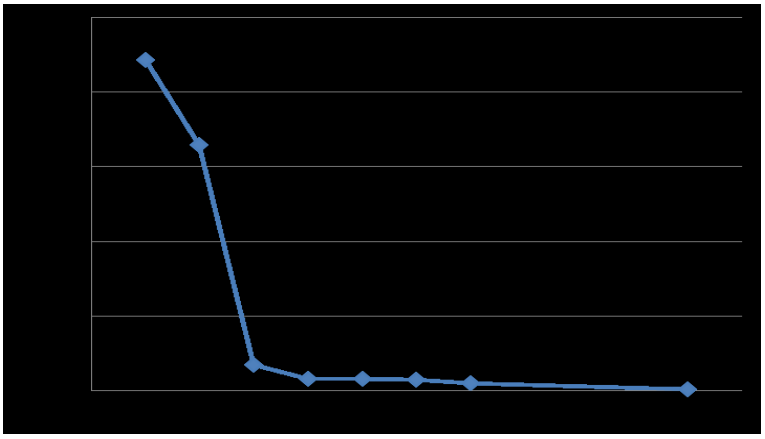
Fig. 6. Resulting histogram (blue color) – the estimation based on: a) the $1^{st} \div 6^{th}$ moments of distribution (*K* = 6; ErrChi2based = 8.01289e-004), b) the $1^{st} \div 7^{th}$ moments of distribution (*K* = 7; ErrChi2based = 7.25156e-004)



a) b)

Rys. 7. Wynikowy histogram (kolor niebieski) – przybliżenie sporządzone na podstawie: a) momentów rozkładu rzędów 1. do 8. (*K* = 8; ErrChi2based = 5,2024e-004), b) momentów rozkładu rzędów od 1. do 12. (*K* = 12; ErrChi2based = 1,1019e-004)

Fig. 7. Resulting histogram (blue color) – the estimation based on: a) the $1^{st} \div 8^{th}$ moments of distribution (*K* = 8; ErrChi2based = 5.2024e-004), b) the $1^{st} \div 12^{th}$ moments of distribution (*K* = 12; ErrChi2based = 1.1019e-004)



Rys. 8. ErrChi2based (K) – zależność pomiędzy błędem rekonstrukcji rozkładu a maksymalnym rzędem momentów rozkładu

Fig. 8. ErrChi2based (K) – dependency between the error of reconstruction the distribution and the maximum of order of the moments

3.3. Wyniki realizacji eksperymentów dla dodatkowych przykładów

Realizacja eksperymentów opisanych powyżej oraz innych, dodatkowych – superpozycja kilku rozkładów Gaussa, tj. od 3 do 5 klastrów gaussowskich – prowadzi do wniosku, że zadowalające wyniki estymacji rozkładu (spełnienie założonego warunku: $\mathrm{ErrChi2based} \le 0{,}001$) można uzyskać, opierając się na znajomości momentów rozkładu rzędu od 1. do ok. 10. ($K \approx 10$).

Eksperymenty zostały wykonane z użyciem programu Matlab; minimum funkcji F zostało wyznaczone za pomocą *fmincon* [4]. Odpowiednie wywołanie jest następujące:

```
…
% opcje dla fmincon - procedury szukania minimum
opt = optimset ('MaxFunEvals' , 100000, 'TolFun', 1e-6,
      'TolX', 1e-6, 'MaxIter' ,10000 , 'Algorithm', 'interior-point');

% minus_entro – minimalizowana funkcja F (minus entropia rozkładu)
% A, M, lb, ub – określenie parametrów ograniczeń
% A, M - macierz i wektor określone wzorem 5
% lb, ub -  wektory zer i jedynek ograniczające dziedzinę pi – wzór 6
% p0 – inicjalna wartość wektora prawdopodobieństw, „punkt” startu fminunc

% p_out – szukany wektor prawdopodobieństw dla histogramu wynikowego

[p_out,fval, exitflag, output]
         = fmincon(@minus_entro, p0,[],[], A , M, lb, ub,[], opt);
output.message
…
```

Realizacja skryptu Matlaba (zawierającego m.in. ww. fragment kodu wywołania *fmincon* oraz wyznaczającego macierz A) tworzącego wartości histogramu wynikowego (o rozdzielczości $N = 50$), czyli obliczająca wektor *p_out* dla $K = 10$ na komputerze z procesorem Intel Core Duo T9600 @ 2.80 GHz, zajęła zaledwie średnio ok. 0,51 s. W ramach realizacji *fmincon* wykonano 65 iteracji algorytmu wyszukiwania minimum (algorytm „inte-

rior-point") i 3444 wywołania funkcji *minus_ento*. Krótkie czasy realizacji ww. programu pozwalają pozytywnie myśleć o praktycznym zastosowaniu omawianej metody.

4. Zaproponowane przykłady metod predykcji wartości momentu rozkładu w chwili o indeksie $t + 1$

Śledzenie zmian rozkładu prawdopodobieństwa w czasie można zrealizować przez śledzenie ewolucji wartości wybranych parametrycznych estymatorów rozkładu. Przykładowo, dysponując wartościami momentów rozkładu w chwilach poprzednich, możemy przewidzieć wartości tych momentów w kolejnej chwili, w przyszłości.

Zagadnienie przewidywania przyszłej, nieznanej wartości momentu można sprowadzić do problemu predykcji ciągów czasowych (inaczej predykcji szeregów czasowych). Problematyka predykcji jest szeroko omawiana i związana jest m.in. z zagadnieniem identyfikacji modelu układów (statycznych/dynamicznych, liniowych/nieliniowych) [10, 11].

W ramach niniejszego artykułu przedstawiono przykładowe dwa wybrane podejścia do zagadnienia predykcji: jedno, wykorzystujące model autonomicznego liniowego dyskretnego układu dynamicznego (układ z dyskretnym czasem) z zakłóceniami, opisanego przez liniowe równanie różnicowe, drugie, oparte na nieliniowym równaniu różnicowym, wykorzystującym sieć neuronową typu RBF (ang. *Radial Basis Function Network*). Oczywiście oba prezentowane podejścia to zaledwie przykłady rozwiązania problemu z szerokiej klasy metod możliwych do zastosowania.

Załóżmy, że celem predykcji jest jeden z wybranych momentów (np. moment rzędu pierwszego – wartość średnia), oznaczony przez m, którego 5 znanych wartości w poprzednich chwilach o indeksach $t-4, t-3, \ldots, t$ wynosi odpowiednio:

$$m = [1 \;\; 1{,}3 \;\; 1{,}5 \;\; 1{,}6 \;\; 1{,}65]. \tag{13}$$

Celem metod zaprezentowanych w poniższych podrozdziałach jest estymacja wartości m_{t+1} w chwili $t+1$. Wybór podejścia (np. jednego z poniższych) zależy od uzyskanej eksperymentalnie szacowanej wartości błędu predykcji i będzie zapewne uzależniony od specyfiki zmian wartości składowych wektora momentów w ramach konkretnego zastosowania.

4.1. Predykcja z wykorzystaniem modelu dyskretnego liniowego układu dynamicznego

Model pozwalający na przewidywanie wartości m w chwili o indeksie $t + 1$ można zbudować w ramach zadania identyfikacji liniowego modelu dynamicznego typu AR (ang. *autoregressive model*), czyli modelu autoregresyjnego L-tego rzędu [8]:

$$m_t + a_1 m_{t-1} + .. + a_L m_{t-L} = e_t, \tag{14}$$

gdzie: e_t – wartość sygnału zakłócenia w chwili t, impuls tzw. szumu białego, a_1, ... , a_L = const to szukane stałe.

Zależność 14 można również zapisać w postaci:

$$A(q)m_t = e_t, \quad A = 1 + aq^{-1} + .. + a_L q^{-L}, \tag{15}$$

gdzie q^{-1} jest operatorem przesunięcia.

Dla 5-elementowego ciągu wartości m, danego wzorem (13), za pomocą programu Matlab zrealizowano zadanie identyfikacji linowego modelu układu dynamicznego [7], określonego równaniem różnicowym stopnia 1. lub 2. (dopuszczalny maksymalny stopień równy dwa wynika z ograniczenia rozmiaru danych; tutaj 5 – rozmiar wektora m):

```
m = [ 1  1.3  1.5  1.6 1.65];
model1 = arx (m, 1)
model1 = arx (m, 2)
```

W rezultacie otrzymano następujące dwa zestawy danych wynikowych, określających model i wartość końcową błędu predykcji (ang. *Final Prediction Error* – FPE) [5] odpowiednio dla *modelu 1* i *modelu 2*:

```
Discrete-time IDPOLY model: A(q)y(t) = e(t)
A(q) = 1 - 1.105 q^-1
Estimated using ARX from data set m
Loss function 0.0118573 and FPE 0.0166003

Discrete-time IDPOLY model: A(q)y(t) = e(t)
A(q) = 1 - 1.706 q^-1 + 0.7253 q^-2
Estimated using ARX from data set m
Loss function 7.70667e-005 and FPE 0.00013872
```

Z powodu mniejszej wartości FPE wybrany zostaje *model 1*, któremu odpowiada równanie różnicowe:

$$m_t - 1{,}706\, m_{t-1} + 0{,}7253\, m_{t-2} = 0, \tag{16}$$

przy założeniu braku zakłócenia e_t.

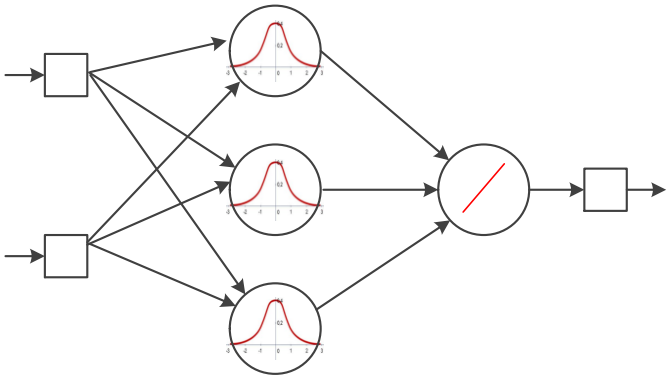
Na podstawie równania (16) można znaleźć ekstrapolowaną wartość m_{t+1} następująco:

$$m_{t+1} = 1{,}706\, m_t - 0{,}7253\, m_{t-1} = 1{,}706 \cdot 1{,}5 - 0{,}7253 \cdot 1{,}65 = 1{,}72695. \tag{17}$$

4.2. Predykcja w wykorzystaniem modelu nieliniowego, opartego na sieci neuronowej typu RBF

Sieci neuronowe typu RBF są często wykorzystywane w aproksymacji funkcji. W sieciach RBF w roli funkcji aktywacji neuronów warstw ukrytych stosuje się tzw. funkcję radialną, której wartości zależą wyłącznie od odległości od jednego wybranego punktu zwanego centrum. Na rys. 9 przedstawiono przykład sieci RBF z 2 neuronami w warstwie wejściowej, 3 neuronami w pierwszej warstwie ukrytej (z radialnymi funkcjami aktywacji neuro-

nów), 1 neuronem w drugiej warstwie ukrytej (z liniową funkcją aktywacji neuronu) i 1 neuronem wyjściowym.



Rys. 9. Przykładowa struktura prostej sieci neuronowej typu RBF
Fig. 9. Sample structure of neural radial basis function neural network

Predykcja wartości m_{t+1} zostanie zrealizowana z wykorzystaniem wyrażenia rekurencyjnego:

$$m_{t+1} = \varphi(m_t, m_{t-1}, ..., m_{t-L}) \tag{18}$$

dla $L = 0, 1, \ldots$, gdzie φ jest $(L + 1)$-argumentową funkcją, której wartość jest określana przez wartość wyjścia $(L + 1)$-wejściowej sieci neuronowej.

Rząd nieliniowego równania różnicowego (określonego formułą (18)), czyli struktura sieci (w tym liczba wejść), będzie określona na drodze minimalizacji uśrednionego błędu standardowego MERR w ramach procedury weryfikacji modelu za pomocą jednoelementowego zbioru testowego (ang. *Leave-one-out*). Wyniki tej weryfikacji zostały przedstawione poniżej.

Dla $L = 0$ równanie (18) ma postać $m_{t+1} = \varphi(m_t)$, a 4-elementowy wektor wejściowy P i 4-elementowy wektor wyjściowy T wynoszą odpowiednio:

$$P = [1 \quad 1{,}3 \quad 1{,}5 \quad 1{,}6], \; T = [1{,}3 \quad 1{,}5 \quad 1{,}6 \quad 1{,}65]. \tag{19}$$

Zastosowanie metody *Leave-one-out* (4 iteracje wyboru elementu testującego) pozwala na wyznaczenie uśrednionego błędu standardowego MERR ≈ 0,04151.

Dla $L = 1$ równanie (18) ma postać $m_{t+1} = \varphi(m_t, m_{t-1})$, a 3-kolumnowa macierz wejściowa P i 3-elementowy wektor wyjściowy T wynoszą odpowiednio:

$$P = \begin{bmatrix} 1 & 1{,}3 & 1{,}5 \\ 1{,}3 & 1{,}5 & 1{,}6 \end{bmatrix}, \; T = [1{,}5 \quad 1{,}6 \quad 1{,}65]. \tag{20}$$

Zastosowanie metody *Leave-one-out* (3 iteracje wyboru elementu testującego) pozwala na wyznaczenie uśrednionego błędu standardowego MERR ≈ 0,10726.

Dla $L = 2$ równanie (18) ma postać $m_{t+1} = \varphi(m_t, m_{t-1}, m_{t-2})$, a 2-kolumnowa macierz wejściowa P i 2-elementowy wektor wyjściowy T wynoszą odpowiednio:

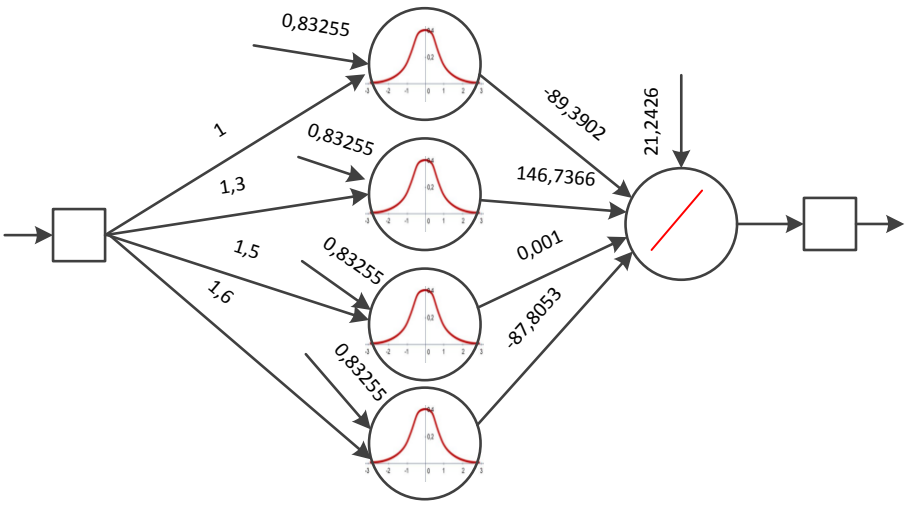
$$P = \begin{bmatrix} 1 & 1{,}3 \\ 1{,}3 & 1{,}5 \\ 1{,}5 & 1{,}61 \end{bmatrix},\ T = [1{,}6 \quad 1{,}65]\ . \tag{21}$$

Zastosowanie metody *Leave-one-out* (2 iteracje wyboru elementu testującego) pozwala na wyznaczenie uśrednionego błędu standardowego MERR ≈ 0,15767.

Biorąc pod uwagę najmniejszą wartość MERR, wybrano model rzędu pierwszego ($L + 1 = 1$), tj. $m_{t+1} = \varphi(m_t)$. Odpowiednią sieć neuronową, uzyskaną przez wykonanie poleceń Matlaba [6]:

```
P = [1    1.3  1.5 1.6]
T = [1.3  1.5  1.6 1.65]
net = newrbe (P, T)
```

przedstawiono na rys. 10.



Rys. 10. Wynikowa sieć neuronowa przeznaczona do predykcji wartości momentu rozkładu w chwili $t + 1$

Fig. 10. Resulting neural network for predicting the moment of distribution in the (t +1)-th moment of time

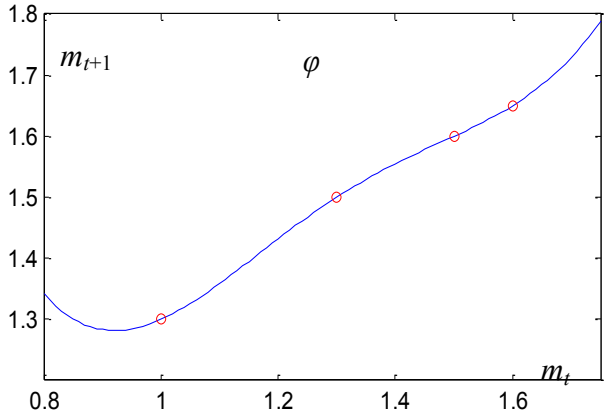
Uzyskaną nieliniową funkcję przejścia φ dla $m_t \in [0{,}75,\ 1{,}75]$ pokazuje rys. 11 (dodatkowo czerwonymi kółkami zaznaczono wartości elementów wektora uczącego).

Ostatecznie szukaną wartość:

$$m_{t+1} \approx 1{,}684 \tag{22}$$

uzyskano z użyciem zbudowanej sieci neuronowej w następujący sposób:

```
m_te_plus_1 = sim(net, 1.65).
```



Rys. 11. Uzyskana nieliniowa funkcja przejścia, pozwalająca na predykcję wartości momentu rozkładu w chwili t +1 na podstawie znanej wartości momentu rozkładu w chwili t

Fig. 11. Resulting non-linear transfer function for prediction of value of distribution moment in the (t + 1)-th moment of time basing on the known value of distribution moment in the t-th moment of time

5. Ekstrapolacja wartości momentu rozkładu po chwili τ_t – interpolacja w dowolnym momencie czasu $\tau \in (\tau_t, \tau_{t+1})$

Przy założeniu, że znana jest wartość pewnego momentu rozkładu w chwilach τ_t, τ_{t-1}, τ_{t-2} ,… (czasy ostatnich aktualizacji statystyki) oraz znana jest oszacowana wartość tego momentu w chwili τ_{t+1} (uzyskana w ramach procedury predykcji opisanej w rozdziale 4), stosując interpolację, można oszacować wartość momentu w dowolnej chwili czasu $\tau \in (\tau_t, \tau_{t+1})$.

Przykładowo załóżmy, że znany jest wektor $\boldsymbol{M_t}$ wartości pewnego momentu rozkładu w pewnych chwilach czasowych, tzn.:

- wartość dokładna w teraźniejszości (indeks t),
- wartości dokładne w przeszłości (indeksy: $t - 1$, $t - 2$,…),
- wartość estymowana w przyszłości (indeks $t + 1$),

$$\boldsymbol{M_t} = [m_{t-4},...,m_t,\hat{m}_{t+1}] = [m(\tau_{t-4}),...,m(\tau_t),m(\tau_{t+1})] = [1\ \ 1{,}3\ \ 1{,}5\ \ 1{,}6\ \ 1{,}65\ \ 1{,}684]. \quad (23)$$

We wzorze (23) symbol $m(\tau)$ oznacza nieznaną funkcję z ciągłą dziedziną argumentów τ o znanych wartościach $m(\tau_j) = m_j$ dla $j = t - 4, \ldots, t, t + 1$.

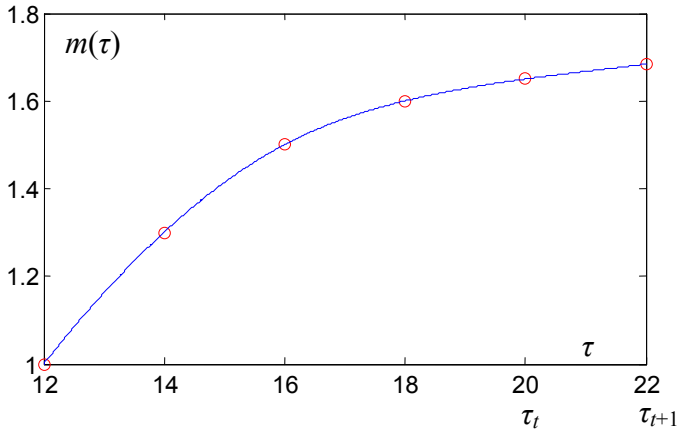
Jeżeli $t = 8$ i $\tau_t = \tau_8 = 20$ oraz

$$\forall_j \tau_{j+1} - \tau_j = \Delta\tau = 2, \quad (24)$$

wówczas wektor momentów czasowych będzie wynosić:

$$\boldsymbol{T} = [\tau_{t-4},...,\tau_t,\tau_{t+1}] = [12,...,18,20,22]. \quad (25)$$

Stosując interpolację, na podstawie wektorów $\boldsymbol{T}$ i $\boldsymbol{M_t}$ można skonstruować funkcję $m(\tau)$, której przebieg dla $\tau \in [12, 22]$ pokazano na rys. 12. W omawianym przykładzie zastosowano interpolację metodą funkcji sklejanych stopnia 3. (ang. *cubic spline interpolation*) [9] (dodatkowo czerwonymi kółkami zaznaczono węzły interpolacji).



Rys. 12. Wynik interpolacji – funkcja $m(\tau)$ skonstruowana na podstawie wektorów T = [12, 14,..., 22] i M_t = [1 1,3 1,5 1,6 1,65 1,684]
Fig. 12. Result of interpolation – function $m(\tau)$ based on T = [12, 14, .., 22] and M_t = [1 1.3 1.5 1.6 1.65 1.684]

Oczywiście celem przedstawionych działań jest interpolacja wartości m w dowolnie wybranej chwili $\tau \in (\tau_t, \tau_{t+1}) = (20, 22)$. Przykładowo dla $\tau = 21$ w wyniku realizacji programu:

```
T = 12:2:22;
MT = [1  1.3  1.5  1.6 1.65 1.684];
tau = 21 ;
m_tau = interp1(T, MT, tau,'spline')
```

uzyskano wartość:

$$m(\tau) = m(21) \approx 1{,}667. \tag{26}$$

6. Opis metody

Rozdział szczegółowo opisuje etapy realizacji metody ekstrapolacji rozkładu wartości atrybutu. W podrozdziale 6.1 opisano czynności przygotowawcze związane ze „strojeniem parametrów metody", tj. m.in. wyznaczenie maksymalnego rzędu uwzględnianych momentów rozkładu oraz wyznaczenie modeli predykcji dla każdego z momentów.

6.1. Czynności dodatkowe

Przy każdorazowej aktualizacji statystyki (w chwili o indeksie t, czyli w momencie czasowym τ_t) następuje:

- usunięcie (jeśli istnieje) „poprzedniego" histogramu *equi-width* $\{(x_i, p_i(\tau_{t-1})\}$ (utworzonego w chwili o indeksie $t-1$),
- zbudowanie „nowego" histogramu *equi-width* $\{(x_i, p_i(\tau_t)\}$ (utworzonego w chwili o indeksie t) na podstawie aktualnej zawartości atrybutu X w bazie danych w momencie czasowym τ_t,
- utworzenie kolejnego wektora wartości momentów rozkładu, tj.:

$$\boldsymbol{m}(\tau_t) = \begin{bmatrix} m^{(1)}(\tau_t) \\ \dots \\ m^{(K)}(\tau_t) \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} x_i p_i(\tau_t) \\ \dots \\ \sum_{i=1}^{N} x_i^K p_i(\tau_t) \end{bmatrix}. \tag{27}$$

Przy pierwszym tworzeniu statystyki (moment czasowy $\tau_0 = 0$), po zbudowaniu histogramu, następuje oszacowanie parametru K (maksymalny uwzględniany rząd momentów rozkładu) z wykorzystaniem kryterium $\mathrm{ErrChi2based} \le \varepsilon$, gdzie ε to zakładany maksymalny próg błędu estymacji histogramu docelowego (wzór (9)).

Jeśli liczba aktualizacji statystyk osiąga zadany próg L (L – mała liczba całkowita, np. 5), tzn. określony jest ciąg wektorów momentów rozkładu:

$$\boldsymbol{m}(\tau_0), \boldsymbol{m}(\tau_1),\dots, \boldsymbol{m}(\tau_L) \tag{28}$$

w chwilach $\tau_0, \tau_1,\dots, \tau_L$, to następuje określenie modelu predykcji (rozdział 4) dla każdego momentu z osobna. Zakłada się, że modele predykcji φ_r, wyznaczone na podstawie ciągów $m^{(r)}(\tau_0), m^{(r)}(\tau_1),\dots, m^{(r)}(\tau_L)$, czyli dla poszczególnych momentów rozkładu określonego rzędu r, mogą być różne zarówno ilościowo (tzw. różny rząd modelu $R_r \le L$ czy różne wartości parametrów modelu), jak i jakościowo (różny rodzaj modelu).

6.2. Zasadniczy algorytm ekstrapolacji rozkładu

Przedstawiona poniżej procedura pozwala na wykonanie ekstrapolacji rozkładu dla dowolnej chwili τ przy uwzględnieniu $\tau_t < \tau < \tau_{t+1}$, gdzie τ_t to moment czasowy ostatniej zrealizowanej aktualizacji statystyki, a τ_{t+1} to zakładany moment czasowy następnej aktualizacji statystyki, w przyszłości.

Zakłada się, że przed uruchomieniem procedury ekstrapolacji dane są:

- histogram *equi width* $\{(x_i, p_i(\tau_t)\}$,
- modele predykcji momentów rozkładu φ_r dla $r = 1...K$,
- wartości momentów rozkładu w poprzednich chwilach czasowych: $m^{(r)}(\tau_{t-R_r+1}), m^{(r)}(\tau_{t-R_r+2}),..., m^{(r)}(\tau_t)$ dla $r = 1...K$, gdzie R_r to liczba chwil czasowych, określona przez rząd modelu predykcji momentu r-tego rzędu.

Procedura ekstrapolacji rozkładu na chwilę τ ($\tau_t < \tau < \tau_{t+1}$) zakłada realizację następujących czynności:

1. predykcję wektora momentów rozkładu w przyszłej chwili τ_{t+1}, tzn. wyznaczenie $\hat{\boldsymbol{m}}(\tau_{t+1})$ na podstawie modeli φ_r dla $r = 1\ldots K$ oraz $\boldsymbol{m}(\tau_t),\ldots, \boldsymbol{m}(\tau_{t-Rr+1})$ (rozdział 4),
2. interpolację $\hat{\boldsymbol{m}}(\tau)$ na podstawie $\hat{\boldsymbol{m}}(\tau_{t+1})$, $\boldsymbol{m}(\tau_t),\ldots, \boldsymbol{m}(\tau_{t-Rr+1})$ (rozdział 5),
3. wyznaczenie docelowego histogramu $\{(x_i, p_i(\tau)\}$, tzn. obliczenie wartości $p_i(\tau)$ przez minimalizację F – „funkcji minus entropii" (rozdział 2), przy ograniczeniach zbudowanych na podstawie wektora $\hat{\boldsymbol{m}}(\tau)$ (zawartego w wektorze $\boldsymbol{M}$ ze wzoru (5)); zakłada się, że wartościami startowymi algorytmu minimalizacji są prawdopodobieństwa z histogramu z ostatniej aktualizacji statystyki, tj. z chwili τ_t (tzn. wektor *p0* w listingu z rozdziału 0 jest inicjowany wartościami $p_i(\tau_t)$).

7. Podsumowanie

Artykuł dotyczy problemu reprezentacji zmiennego w czasie rozkładu wartości atrybutów w kontekście wyznaczania selektywności zapytań opartej na takiej reprezentacji. Artykuł dotyczy sytuacji, w której aktualizacja statystyk – tworzenie histogramów reprezentujących rozkład – jest czasochłonna i może być rzadko wykonywana. W czasie pomiędzy aktualizacjami możliwa jest zmiana danych, a tym samym staje się możliwa zmiana rozkładu. W takich przypadkach użyteczny mógłby być zaproponowany mechanizm programowy, pozwalający na ekstrapolację reprezentacji w chwilach pomiędzy aktualizacjami. Wówczas selektywność mogłaby być wyznaczana na podstawie ekstrapolowanej postaci rozkładu, a nie z wykorzystaniem mniej aktualnej postaci rozkładu, pochodzącej z ostatniej aktualizacji statystyk.

W opracowaniu zaproponowano metodę, w której śledzi się ewolucję momentów rozkładu, a następnie wyznacza się ekstrapolowany rozkład, stosując zasadę maksimum entropii informacyjnej szukanego rozkładu, przy ograniczeniach wynikających z przewidywanych wartości momentów rozkładu w przyszłości.

Dalsze prace mogą się koncertować na pogłębionej, ilościowej weryfikacji metody (np. określenie związku pomiędzy maksymalnym rzędem uwzględnianych momentów, różną postacią rozkładów oraz różną rozdzielczością histogramu wynikowego dla zadanego błędu estymacji).

Innym kierunkiem rozwoju metody może być jakościowa modyfikacja (i weryfikacja) zaproponowanej metody w alternatywnym wariancie, w którym zamiast śledzenia zmian momentów rozkładu rzędu $1\ldots K$ można byłoby zastosować śledzenie kwantyli rozkładu, czyli ekstrapolację rozkładu na podstawie ewolucji kwantyli K-tego rzędu. Takie podejście pozwa-

la na obsługę szerszej klasy rozkładów, ponieważ momenty rozkładu mogą nie istnieć dla pewnych specyficznych rozkładów, a kwantyle istnieją zawsze.

Dalsze prace będą mogły się koncentrować na praktycznym zastosowaniu metody, tzn. implementacji omawianego podejścia w ramach konkretnego SZBD. Z pewnością będzie się dało zaimplementować (z użyciem języków Java i PL/SQL) omawianą metodę jako rozszerzenie SZBD Oracle, wykorzystując moduł *ODCI Stats* [14] do rozszerzenia funkcjonalności tworzenia statystyk i optymalizatora zapytań, tak jak to miało miejsce w zastosowaniach [12] i [13].

BIBLIOGRAFIA

1. Jaynes E. T.: Papers on Probability, Statistics, and Statistical Physics. Springer, 1989.
2. Buck B., Macaulay V. A.: Maximum entropy in action: a collection of expository essays. Clarendon Press, 1991.
3. Saad T.: The Maximum Entropy Method for Reconstructing Density Distributions, 2013, http://www.tsaad.net/docs/tsaad-maximum-entropy-method.pdf.
4. Find minimum of constrained nonlinear multivariable function – MATLAB, 2013, http://www.mathworks.com/help/optim/ug/fmincon.html;jsessionid=efda3f7c6d73ec5a5ed6bd50605e?s_tid=doc_12b.
5. Akaike Final Prediction Error for estimated model – MATLAB, 2013, http://www.mathworks.com/help/ident/ref/fpe.html.
6. Design exact radial basis network – MATLAB, 2013, http://www.mathworks.com/help/nnet/ref/newrbe.html.
7. System Identification Toolbox Documentation – MATLAB, 2013, http://www.mathworks.com/help/ident/index.html#linear-model-identification.
8. Niederliński A.: Systemy komputerowe automatyki przemysłowej. Zastosowania. Tom 2. WNT, Warszawa 1985.
9. 1-D data interpolation – MATLAB, 2013, http://www.mathworks.com/help/matlab/ref/interp1.html.
10. Ljung L.: System Identification: Theory for the User. Prentice Hall 1998.
11. Haber R., Keviczky L.: Nonlinear System Identification – Input-Output Modeling Approach. Springer 1999.
12. Augustyn D. R.: Applying advanced methods of query selectivity estimation in Oracle DBMS. Advances in Soft Computing. Man-Machine Interactions. Springer-Verlag, Berlin-Heidelberg 2009, s. 585÷593.

13. Augustyn D. R.: Zastosowanie sieci Bayesa w szacowaniu selektywności zapytań w optymalizatorze zapytań serwera bazy danych Oracle. Studia Informatica, Vol. 32, No. 1A (94), Gliwice 2011, s. 25÷42.
14. Oracle 10g. Using extensible optimizer, 2010, http://download.oracle.com/docs/cd/B14117 01/appdev.101/b10800/dciextopt.htm.

Wpłynęło do Redakcji 16 stycznia 2013 r.

Abstract

Query optimization is a process which leads to obtain the best query execution method, so-called the execution plan. To find the optimal execution method a selectivity parameter is needed. It enables to estimate a size of data which satisfying a selection condition of analyzed query. Obtaining the query selectivity requires a non-parametric estimator of attribute value distribution, e.g. a equi-width histogram. Histograms are produced during update statistics process. For large databases update statistics process is performed rather seldom, e.g. only during time of low workload of a system. This results that histograms do not describe actual data distribution.

To obtain a more accurate histogram, a prediction mechanism should be introduced. This results obtaining a more accurate estimation of selectivity. The method of extrapolation of attribute value distribution is proposed in this paper. This method tracks the evolution of distribution moments in the past. Using known previous values of distribution moments, the method predicts future values of moments of the extrapolated distribution. Finally, it uses the maximum entropy principle for obtaining the extrapolated distribution subject to the predicted values of the distribution moments.

Adres

Dariusz Rafał AUGUSTYN: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16, 44-100 Gliwice, Polska, draugustyn@polsl.pl.