ЗНАЧАЈ ВЕЛИЧИНЕ БИНОВА ХИСТОГРАМА У АНАЛИЗИ ПОДАТАКА

Вишња Огњеновић[1] Владимир Бртка [2] Ивана Берковић [3] Елеонора Бртка [4]

Резиме: Хистограми дају графичку представу посматраних вредности. За континуалне податке бинови хистограма су у директној вези са дискретизацијом података, која је полазна тачка у препроцесирању континуалних података. У раду је приказан значај величине бинова хистограма као важан део препроцесирања података. Показани су начини дефинисања бинова хистограма. Могуће је користити одређени алгоритам за дискретизацију података у циљу израчунавања минималног броја бинова. Рад је поткрепљен примером.

Кључне речи: хистограм, бин, дискретизација, препроцесирање података, анализа података

SIGNIFICANCE OF HISTOGRAM BINS SIZE IN DATA ANALYSIS

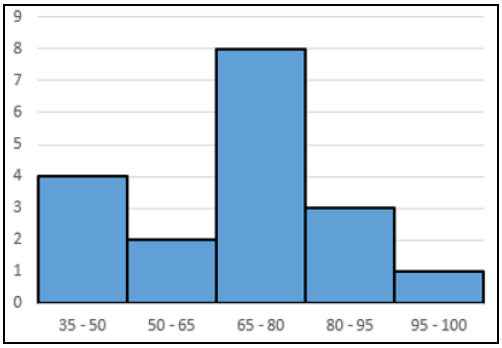
Abstract: Histograms give a graphical representation of the observed values. For continuous data, histogram bins are directly related to data discretization, which is the starting point in preprocessing of continuous data. The significance of histogram bins size as an important part of data preprocessing is presented in the paper. Ways to define bin histograms are shown. It is possible to use a particular algorithm to discretize data in order to calculate the minimum number of bins. The work is supported by an example.

Key words: histogram, bin, discretization, data preprocessing, data analysis

1. УВОД

Хистограми дају графичку представу посматраних вредности. У координатном систему са две осе, на апциси су означене вредности интервала, а на ординати број појављивања свих вредности из одређеног интервала. Таква фреквентност подразумева да су све почетне вредности најпре груписане у интервале по неком правилу, формули или законитости. Најчешће су интервали исте ширине, али није обавезно. С обзиром на велико коришћење хистограма у разним софтверима где је углавном иста ширина свих интервала, у овом раду ће се разматрати само интервали исте ширине.

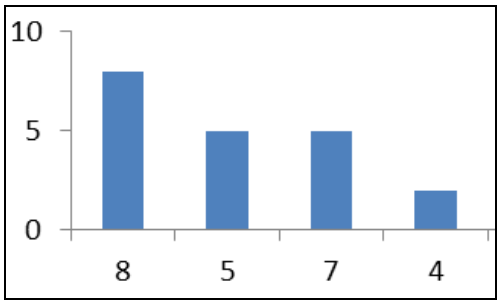
Након дељења почетног интервала на апциси на подинтервале исте ширине, бинови се одређују као правоугаоници чија је једна страница подинтервал, а друга станица фреквентност која се очитава на ординати, као што је приказано на Слици 1.



Слика 1 – Хистограм дискретних вредности

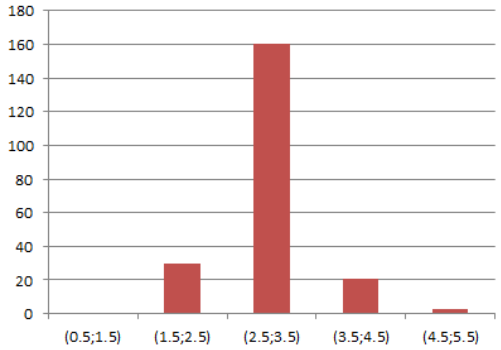
[1] Доц. др, Технички факутет „Михајло Пупин“, Ђуре Ђаковића бб, Зрењанин, visnjao@tfzr.uns.ac.rs
[2] Проф. др, Технички факутет „Михајло Пупин“, Ђуре Ђаковића бб, Зрењанин, brtkav@gmail.com
[3] Проф. др, Технички факутет „Михајло Пупин“, Ђуре Ђаковића бб, Зрењанин, berkovic@tfzr.uns.ac.rs
[4] Доц. др, Технички факутет „Михајло Пупин“, Ђуре Ђаковића бб, Зрењанин, eleonorabrtka@gmail.com

За дискретне вредности, хистограми који се генеришу разним софтверима често не групишу више дискретних вредности, већ се бин формира над сваким бројем, као што се види на Слици 2.



Слика 2 – Хистограм дискретних вредности

За континуалне вредности дискретизација података је неопходна. На Слици 3 је дат приказ хистограма континуалних вредности.



Слика 3 – Хистограм континуалних вредности

Битна разлика је у томе што код континуалних података доња страница бина представља интервал.

2. ПРАВИЛА И ХЕУРИСТИКЕ ЗА ОДРЕЂИВАЊЕ ОПТИМАЛНИХ БИНОВА

Постоји више правила и хеуристика за одређивање одговарајућих бинова, односно броја бинова у односу на домен вредности података. Значајан број правила је повезан са анализом функције расподеле која је најсличнија посматраном хистограму.

2.1. Веза хистограма са функцијом густине одговарајуће расподеле

Са становишта вероватноће, један од најчешћих начина за описивање расподеле вероватноће је одређивање његове функције густине вероватноће (Gholamy, Kreinovich, 2017), као што је приказано формулом (1).

$$\rho(x) \overset{def}{=} \frac{dp}{dx} = \lim_{h \to 0} \frac{\Pr ob(X \in [x, x+h])}{h} \tag{1}$$

У конкретним ситуацијама, све што знамо о дистрибуцији вероватноће је узорак података који одговара овој дистрибуцији. У узорку је могуће за неку малу вредност од x проценити вредност функције. По дефиницији, вероватноћа догађаја је једнака лимесу као у формули (2),

$$\Pr ob(X \in [x, x+h]) = \lim_{n \to \infty} \frac{n([x, x+h])}{n} \tag{2}$$

где је *n* укупан број тачака података, а $n([x, x+h])$ означава број тачака у интервалу $[x, x+h]$. Због тога, се према (Gholamy, Kreinovich, 2017), као процена за одговарајућу вероватноћу, добија фреквенција догађаја *f* према формули (3).

$$f([x, x+h]) \approx \frac{n([x, x+h])}{n} \tag{3}$$

Очекивање за функцију густине вероватноће $\rho(x)$, који је први описао Karl Pearson у (Pearson, 1895), а односи се на очекивање дато формулом (4) зове се апроксимација хистограма (histogram approximation).

$$\rho(x) \approx \frac{f([x, x+h])}{h} \tag{4}$$

У неким софтверима за приказ хистограма, најприближнија функција густине вероватноће је нацртана преко хистограма, односно преко бинова (EasyFit, 2019).

У складу са начином дефинисања хистограма могуће се анализирати дефиниције бинова. Најједноставнији приказ бинова је помоћу интервала исте ширине. То значи да је интервал свих могућих вредности од *x* подељен на подинтервале $[x_i, x_{i+1}]$.

У ситуацијама када се имају додатне информације о одговарајућој расподели вероватноће, према (Gholamy, Kreinovich, 2017), проблем одређивања бинова се може третирати као проблем оптимизације, док се у већини случајева оптимална величина бинова (h_{opt}) смањује са бројем тачака *n* података као у формули (5).

$$h_{opt} = const \frac{s}{n^{\frac{1}{3}}} \tag{5}$$

где је *s* ширина интервала.

2.2. Правила и хеуристике за одређивање бинова на основу густине функције расподела

Правила и хеуристике за одређивање оптималних бинова су развијене на основу мера сличности са одговарајућом функцијом расподеле, где је посматрано одступање од функције расподеле при промени бинова, попут формуле (4).

Sturger-ово правило (Hyndman, 1995) за одређивање оптималних бинова је дато формулом (6) и оно је оптимално само за нормалну симетричну расподелу.

$$K = 1 + 3.322 \cdot \log_N \tag{6}$$

где је K – број бинова а N – број елемената скупа.

Scott-ово правило је базирано на стандардној девијацији података а формула је дата једначином (7).

$$B = 3.49\sigma n^{-\frac{1}{3}} \tag{7}$$

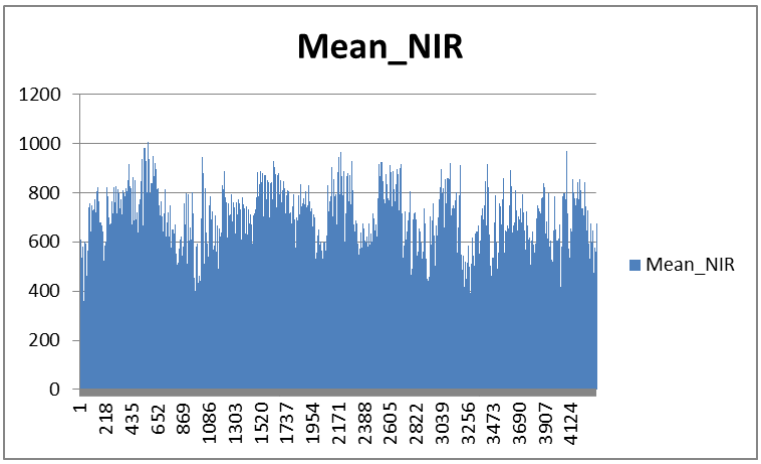
где је *B* – број бинове, σ - стандардна девијација података, а *n* – број елемената скупа**.**

Постоје алгоритми за одређивање оптималне величине бинова базиране на степенастом Бајесовом правилу (He, Meeden, 1997), као и други алгоритми.

2.3. Начини за одређивање бинова на основу дискретизације података

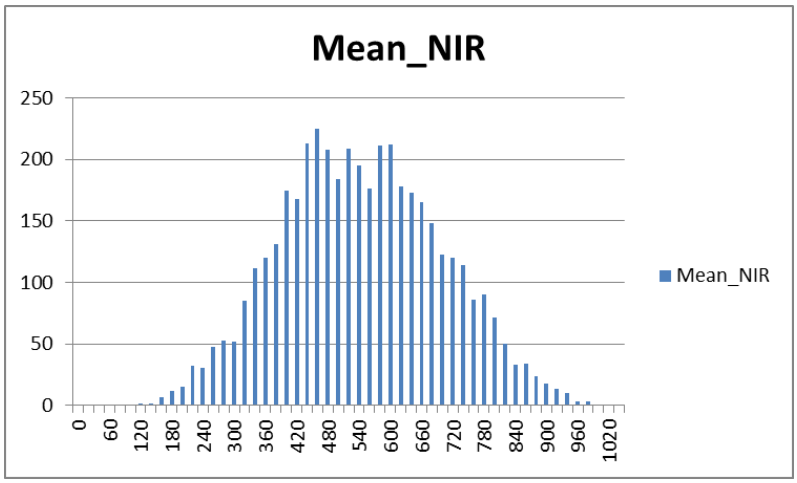
Оно што је посебно занимљиво је то да се при промени величине подинтервала, не само мења број бинова, већ и сам хистограм.

На Слици 4 приказан је хистограм атрибута Mean_NIR, базе Wilt Data Set (Johnson, Tateishi, Hoan, 2013), без дискретизације вредности.



Слика 4 – Хистограм атрибута Mean_NIR без дискретизације

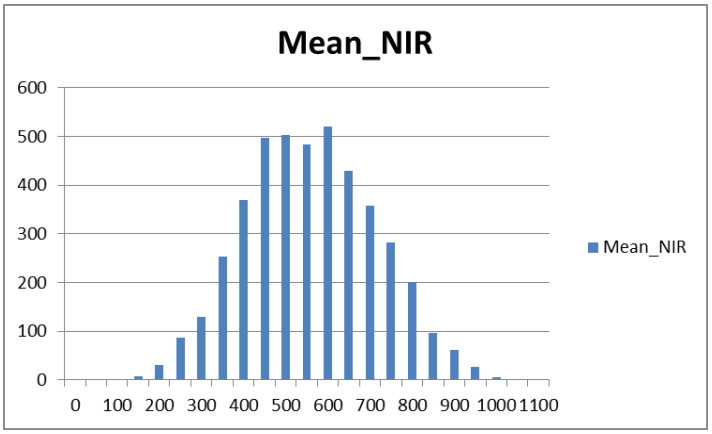
На Слици 5 приказан је хистограм истог атрибута Mean_NIR, базе Wilt Data Set, са урађеном једноставном дискретизацијом типа једнаких интервала ширине 20 као што се види на апциси.



Слика 5 – Хистограм атрибута Mean_NIR са дискретизованим вредностима – ширина бина 20

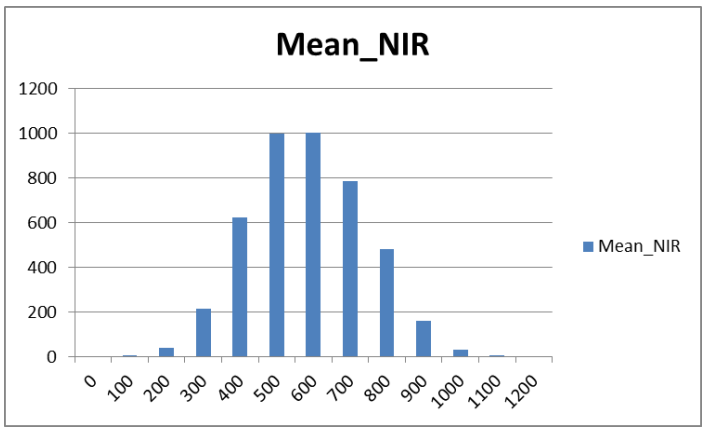
Разлика између хистограма са Слике 4 и Слике 5 је очигледна. На Слици 4 се не назире облик хистограма, док се са Слике 5 види да је у питању нека унимодал расподела.

Ако се смањи број подинтервала, односно ако се изабере „грубља“ дискретизација континуалних вредности са ширином бина 50, онда ће хистограм изгледати као на Слици 6.



Слика 6 – Хистограм атрибута Mean_NIR са дискретизованим вредностима – ширина бина 50

У случају ако се за исти атрибут узме још мањи број бинова, изгубиће се прецизност, али се у овом конкретном случају још увек види да је у питању унимодал расподела (Слика 6).



Слика 6 – Хистограм атрибута Mean_NIR са дискретизованим вредностима – ширина бина 50

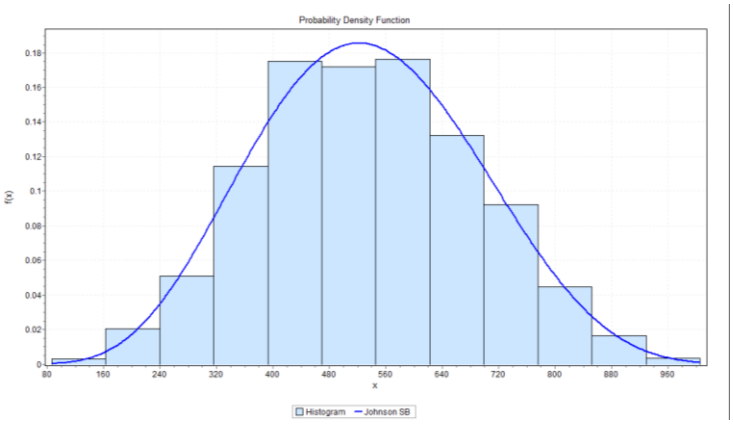
Могуће је одређивати бинове и на основу дискретизованих података по неком конкретном алгоритму (Ognjenović, 2016). Да би сви бинови имали исту ширину, потребно је одредити најмањи подинтервал добијен дискретизацијом података. На тај начин би се сачувале основне вредности добијене дискретизацијом података.

3. ПРИМЕР - ПРЕДЛОГ ПРИМЕНЕ АЛГОРИТМА ЗА ДИСКРЕТИЗАЦИЈУ

За изабрану дискретизацију максималног разликовања података (Nguyen, 2006), примењену на исте податке као у делу 2.3. слика 4, базе Wilt Data Set, а добијену софтвером ROSETTA (Øhrn, 2001), подинтервали атрибута Mean_NIR су одређени следећим тачкама: 2.54304e+007, 3.7212e+007 и 4.53399e+007.

Најмање растојање подинтервала је око 80, што значи да је то довољна минимална ширина бинова. На Слици 7 је приказан хистограм са ширином бинова 80 и нацртаном функцијом густине расподеле.

Решење одређивања минималног броја бинова у складу са алгоритмом максималног разликовања представља лако видљив и препознатљив хистограм.



Слика 7 – Хистограм атрибута Mean_NIR са дискретизованим вредностима – ширина бина 80

4. ЗАКЉУЧАК

У раду је указано на значај одређивања оптималног броја бинова хистограма. Дат је предлог за одређивање минималног броја бинова хистограма у складу са изабраним алгоритмом дискретизације података. На примеру је показано одређивање минималног броја бинова хистограма једначењем ширине бина са најмањом ширином подинтервала добијеног дискретизацијом података.

ЗАХВАЛНИЦА

Овај рад је подржан од стране Министарства образовања и науке Републике Србије у оквиру пројекта TR32044 „Развој софтверских алата за анализу и побољшање пословних процеса“, 2011-2019.

5. ЛИТЕРАТУРА

[1] EasyFit (2019). EasyFit: Distribution Fitting Made Easy, http://www.mathwave.com/easyfit-distribution-fitting.html
[2] Gholamy A., Kreinovich V., (2017), What Is the Optimal Bin Size of a Histogram: An Informal Description, International Mathematical Forum, 2017, Vol. 12, No. 15, pp. 731-736
[3] He K., Meeden G., (1997). Selecting the Number of Bins in a Histogram: A Decision Theoretic Approach, Appeared in Journal of Statistical Planning and Inference,Vol 61(1997),59-59.
[4] Hyndman R. (1995), The problem with Sturges' rule for constructinghistograms, https://robjhyndman.com/papers/sturges.pdf
[5] Johnson, B., Tateishi, R., Hoan, N., (2013), A hybrid pansharpening approach and multiscale object-based image analysis for mapping diseased pine and oak trees. International Journal of Remote Sensing, 34 (20), 6969-6982.
[6] Nguyen H.S. (2006), Approximate boolean reasoning: foundations and applications in data mining, Transactions on rough sets V, 334-506.
[7] Ognjenović V. (2016). Aproksimativna diskretizacija tabelarno organizovanih podataka, doktorska disertacija, Tehnički fakultet „Mihajlo Pupin“ Zrenjanin
[8] Øhrn A. (2001), ROSETTA Technical Reference Manual, Department of Computer and Information Science, Norwegian University of Science and Technology, www.lcb.uu.se/tools/rosetta/materials/manual.pdf
[9] Pearson K. (1895), Controbutions to the mathematical theory of evolution.II. Skewvariation in homogeneous matreial, Philosophical Transactions of the RoaySociety A: Mathematical, Physical, and Engineering Sciences, Vol. 186,pp. 343–414.