PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 176

Taksonomia 18. Klasyfikacja i analiza danych – teoria i zastosowania 2011

Małgorzata Gliwa

Uniwersytet Ekonomiczny w Katowicach

WPŁYW METODY DYSKRETYZACJI NA JAKOŚĆ KLASYFIKACJI

Streszczenie: Główny cel artykułu to porównanie wielkości błędów klasyfikacji modeli dyskryminacyjnych zbudowanych dla zbiorów danych przed dyskretyzacją i po dyskretyzacji. Jako metodę dyskryminacji zastosowano naiwny klasyfikator bayesowski. Modele budowano dla zbiorów danych zarówno przed dyskretyzacją, jak i po dyskretyzacji. Dyskretyzacji dokonano z wykorzystaniem metod bezkontekstowych (dyskretyzacja na równe przedziały i przedziały o równych liczebnościach) i kontekstowych (metoda ChiMerge i minimalizacji entropii). Obliczenia wykonano na podstawie autorskich procedur i funkcji zawartych w pakietach `dprep`, `e1071`, `grDevices`, `infotheo` oraz `car` programu **R**.

Słowa kluczowe: dyskretyzacja zmiennej ciągłej, naiwny klasyfikator bayesowski.

1. Wstęp

Dyskretyzacja zmiennej ciągłej polega na podziale uporządkowanego zbioru wartości danej zmiennej na skończoną liczbę rozłącznych przedziałów. Powstałym przedziałom można przypisać etykiety, otrzymując w ten sposób zmienną nominalną o pewnej liczbie kategorii.

Z jednej strony dyskretyzacja powoduje utratę informacji w zbiorze danych, z drugiej zaś rozwiązuje problem wartości oddalonych oraz braków danych. Inną przesłanką przemawiającą za dyskretyzacją zmiennych są dynamicznie rozwijające się metody analizy obiektów jakościowych (symbolicznych). Obiekty takie charakteryzowane są m.in. przez: zmienną nominalną, listy kategorii czy przedziały liczbowe [Bock, Diday 2000]. Również wśród metod wielowymiarowej analizy danych istnieją takie, np. naiwny klasyfikator bayesowski, dla których zaleca się, aby zbiór danych zwierał obiekty reprezentowane przez zmienne jakościowe. Stosowanie powyższych metod wymusza potrzebę dyskretyzacji zmiennych ciągłych.

Celem niniejszego artykułu jest porównanie jakości klasyfikacji modeli dyskryminacyjnych otrzymanych na podstawie zbiorów danych przed dyskretyzacją i po dyskretyzacji. Jakość klasyfikacji rozumiana jest jako wielkość błędu klasyfika-

cji liczona dla zbioru testowego (im mniejszy błąd, tym lepsza jakość modelu). W przypadku zbiorów zdyskretyzowanych porównane zostaną wielkości błędów klasyfikacji dla różnych metod dyskretyzacji. Do zbudowania modeli dyskryminacyjnych zostanie zastosowany naiwny klasyfikator bayesowski. Badanie empiryczne realizowane będzie za pomocą symulacji komputerowych w programie **R**.

Eksperyment służyć będzie weryfikacji następujących hipotez badawczych:

1. Modele dyskryminacyjne budowane na podstawie zbiorów zdyskretyzowanych charakteryzują się lepszą jakością klasyfikacji niż modele otrzymane dla zbiorów bez dyskretyzacji zmiennych.

2. Lepszą jakością klasyfikacji charakteryzują się modele dyskryminacyjne budowane na podstawie zbiorów dyskretyzowanych za pomocą metod kontekstowych niż modele otrzymane dla zbiorów dyskretyzowanych metodami bezkontekstowymi.

2. Metody dyskretyzacji zmiennej ciągłej zastosowane w badaniu empirycznym

Dougherty, Kohavi i Sahami [1995] wyróżniają następujące podejścia w zagadnieniu dyskretyzacji:

1. Metody bezkontekstowe (*unsupervised*) i kontekstowe (*supervised*). Nazwy tych metod w języku polskim zaproponowane zostały w pracy Gatnara [2000]. Bezkontekstowe metody dyskretyzacji nie uwzględniają informacji o przynależności obiektów do klas (lub wartości innej zmiennej, np. zależnej), w przeciwieństwie do metod kontekstowych, które tę informację uwzględniają.

2. Metody lokalne (*local*) i globalne (*global*). Metody lokalne dokonują dyskretyzacji zmiennej ciągłej w pewnym fragmencie jej dziedziny w zależności od wartości innych zmiennych. Natomiast metody globalne dyskretyzują daną zmienną raz, niezależnie od wartości innych zmiennych.

3. Metody statyczne (*static*) i dynamiczne (*dynamic*). Metody statyczne to te, które poszukują właściwej liczby przedziałów dla każdej zmiennej, niezależnie od wartości pozostałych zmiennych. Metody dynamiczne poszukują punktów podziału dla wszystkich zmiennych ciągłych równocześnie.

Zaprezentowane podejścia nie są rozłączne. Oznacza to, że wśród metod bezkontekstowych i kontekstowych można wskazać zarówno metody lokalne, jak i globalne, statyczne, jak i dynamiczne.

W badaniach empirycznych zastosowano dyskretyzację na przedziały o równej szerokości (`disc.ew`), na przedziały o równej liczebności (`disc.ef`), metodę Chi-Merge (`chiMerge`) i dyskretyzację metodą minimalizacji entropii w oparciu o zasadę minimalnej długości kodu (`disc.mentr`). Dwie pierwsze metody należą do grupy metod bezkontekstowych, dwie następne do grupy metod kontekstowych. W nawiasach podano nazwy funkcji z pakietu `dprep`.

2.1. Dyskretyzacja na przedziały o równej szerokości

Dyskretyzacja na *k* przedziałów o równej szerokości polega na podziale wartości zmiennej ciągłej X na *k* przedziałów o równej rozpiętości *h*, która wyznaczana jest jako:

$$h = \frac{x_{\max} - x_{\min}}{k}, \tag{1}$$

wówczas punkty podziału obliczane są następująco:

$$x_{\min} + \sum_{i=1}^{k-1} ih\,. \tag{2}$$

Liczbę przedziałów użytkownik może wyznaczyć w dowolny sposób lub na podstawie jednej z poniższych formuł:

1. Formuła Sturgesa [1926]:

$$k = \lceil 1 + \log_2 N \rceil. \tag{3}$$

2. Formuła Scotta [1979]: W

$$k = \left\lceil \frac{x_{\max} - x_{\min}}{h} \right\rceil, \tag{4}$$

gdzie: $h = 3{,}49sN^{-\frac{1}{3}}$, s – odchylenie standardowe danej zmiennej.

3. Formuła Freedmana–Diaconisa [1981]:

$$k = \left\lceil \frac{x_{\max} - x_{\min}}{h} \right\rceil, \tag{5}$$

gdzie $h = 2IQN^{-\frac{1}{3}}$, $IQ = Q_3 - Q_1$, Q_1 i Q_3 to odpowiednio kwartyl pierwszy i trzeci.

Symbol $\lceil k \rceil = \min\{x \in Z : k \le x\}$ oznacza najmniejszą liczbę całkowitą nie mniejszą od *k*.

2.2. Dyskretyzacja na przedziały o równej liczebności

W rozważanej metodzie zamiast przedziałów o równej szerokości otrzymuje się przedziały równoliczne. Dokonuje się podziału zbioru wartości zmiennej ciągłej na *k* przedziałów. Liczbę przedziałów wyznacza się w ten sam sposób, jak w przypadku dyskretyzacji na przedziały o równej szerokości. Końce przedziałów ustala się tak, aby w każdym z nich znajdowała się taka sama liczba obserwacji. To oznacza, że dla zbioru *N*-elementowego w każdym przedziale powinno znajdować się N/k obiektów.

2.3. Metoda ChiMerge

Metoda ChiMerge została zaproponowana przez Kerbera [1992]. Według tej metody każda uporządkowana rosnąco wartość danej zmiennej ciągłej jest przydzielana do osobnego przedziału lewostronnie domkniętego. Dolną granicę przedziału stanowi wartość poprzednia, a górną wartość następna. Dalej łączy się sąsiadujące przedziały. Procesem łączenia steruje test niezależności χ^2. W tym celu weryfikuje się hipotezę zerową, mówiącą o braku związku między sąsiednimi przedziałami. Odrzucenie hipotezy zerowej oznacza zależność tychże przedziałów. Zatem mogą one zostać połączone.

Sprawdzianem hipotezy zerowej jest statystyka χ^2:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{m}\frac{\left(n_{ij} - \hat{n}_{ij}\right)^2}{\hat{n}_{ij}}, \tag{6}$$

gdzie: m – liczba klas,

n_{ij} – liczba obiektów z j-tej klasy w i-tym przedziale,

$\hat{n}_{ij} = \dfrac{\sum_{i=1}^{2} n_{ij} \cdot \sum_{j=1}^{m} n_{ij}}{N}$ – tzw. liczebności teoretyczne,

$\sum_{i=1}^{2} n_{ij}$ – liczba obiektów w i-tym przedziale,

$\sum_{j=1}^{m} n_{ij}$ – liczba obiektów w j-tej klasie, N – liczebność zbioru.

Kryterium stopu określone jest przez wartość krytyczną χ^2_α odczytaną z tablic rozkładu χ^2 dla $m - 1$ stopni swobody oraz ustalonego przez użytkownika poziomu istotności α.

Proces łączenia przedziałów przebiega w dwóch etapach i polega na:

1. Obliczeniu wartości statystyki χ^2 dla każdej pary sąsiadujących przedziałów.

2. Połączeniu tej pary sąsiadujących przedziałów, dla której wartość statystyki χ^2 jest najmniejsza.

Przedziały są łączone, dopóki $\chi^2 > \chi^2_\alpha$ [Kerber 1992, s. 124-125].

2.4. Dyskretyzacja metodą minimalizacji entropii

W dyskretyzacji metodą minimalizacji entropii do znalezienia punktów podziałów zbioru wartości dyskretyzowanej zmiennej wykorzystuje się minimalną wartość

funkcji entropii [Catlett 1991; Fayyad, Irani 1993]. Poszukiwana jest wartość *C*, która minimalizuje wyrażenie:

$$E(X,C,S)=\frac{|S_1|}{N}Ent(S_1)+\frac{|S_2|}{N}Ent(S_2), \tag{7}$$

gdzie: $Ent(S)=-\sum_{j=1}^{m}p_j\log_2(p_j)$ – funkcja entropii,

X – zmienna ciągła poddawana dyskretyzacji,

C – punkt podziału,

S – zbiór danych,

$|S_1|$ – liczebność zbioru o wartościach nie większych od *C*, $|S_2|$ – liczebność zbioru o wartościach większych od *C*,

$p_j=\frac{l_j}{N}$ – prawdopodobieństwo przynależności do *j*-tej klasy, l_j to liczebność *j*-tej klasy.

Jeżeli zbiór S zawiera dużo obiektów, to procedura poszukiwania punktu podziału *C*, spośród $N-1$ możliwości, może być bardzo czasochłonna. Czas poszukiwania optymalnego punktu podziału można ograniczyć, wykorzystując twierdzenie udowodnione przez Fayyada i Iraniego [1993], które głosi:

Jeżeli C minimalizuje wartość miary (7), to jest ono punktem granicznym.

Punkt graniczny *C* rozumiany jest jako średnia wartości zmiennej ciągłej obiektów z różnych klas. Zauważyć jednak należy, że w zbiorze S wystąpić mogą dwa obiekty z różnych klas, dla których określona zmienna przyjmuje tę samą wartość. Gatnar [2000] proponuje modyfikację, polegającą na rozważeniu podziału zbioru wartości w punktach rozdzielających przedziały, w których obiekty należą do różnych klas lub z których przynajmniej jeden nie jest homogeniczny.

W podejściu zaproponowanym przez Fayyada i Iraniego [1993] generowane są od razu wszystkie możliwe punkty podziału. Wybór optymalnego punktu podziału odbywa się w oparciu o zasadę minimalnej długości kodu, która stanowi kryterium stopu. Procedura rekurencyjnego podziału zbioru wartości zmiennej *X* zostaje zatrzymana, jeśli spełniona będzie nierówność:

$$Gain(X,C,S)<\frac{\log_2(N-1)+\Delta(X,C,S)}{N}, \tag{8}$$

gdzie: $Gain(X,C,S)=Ent(S)-E(X,C,S)$,

$\Delta(X,C,S)=\log_2(3^k-2)-\left[kEnt(S)-k_1Ent(S_1)-k_2Ent(S_2)\right]$,

k, k_1, k_2 – liczba klas odpowiednio w zbiorze S, S_1, S_2.

3. Badanie empiryczne

3.1. Metoda dyskryminacji zastosowana w badaniu empirycznym

Poszczególne modele dyskryminacyjne budowane będą z zastosowaniem naiwnego klasyfikatora bayesowskiego. Wspomnianą metodę można stosować zarówno w odniesieniu do danych ilościowych, jak i jakościowych. Douherty, Kohavi, Sahami [1995], Hsu, Huang, Wong [2000] oraz Yang, Webb [2001] postulują jednak, aby klasyfikator ten stosować dla danych jakościowych. Wówczas zmienne ciągłe należy zdyskretyzować.

Zadaniem naiwnego klasyfikatora bayesowskiego jest przydzielenie nowej obserwacji $\boldsymbol{x}_i$ do klasy P_j $(j=1,...,J)$, dla której prawdopodobieństwo tego, że obserwacja $\boldsymbol{x}_i$ należy do klasy P_j jest największe [Walesiak, Gatnar 2009, s. 193-194]:

$$\hat{f}(\boldsymbol{x}_i) = P_j, \text{ gdy } p(P_j \mid \boldsymbol{x}_{i\cdot}) = \max_{k=1,...,J} \{p(P_k \mid \boldsymbol{x}_i)\}, \tag{9}$$

gdzie: $p(P_j \mid \boldsymbol{x}_i) = \dfrac{g(\boldsymbol{x}_i \mid P_j) \cdot p(P_j)}{\sum_{k=1}^{J} g(\boldsymbol{x}_i \mid P_k) \cdot p(P_k)}$ – prawdopodobieństwo *a posteriori*[1],

$p(P_j)$ – prawdopodobieństwo *a priori* dla klasy P_j; najczęściej wyznaczane jest jako frakcja obiektów należących do klasy P_j, tj. $p(P_j) = n_j / N$, gdzie $\sum_{j=1}^{J} p(P_j) = 1$.

W metodzie tej zakłada się warunkową niezależność zmiennych $X_1,..., X_m$ dla ustalonej klasy P_j:

$$p(X_1,..., X_m \mid P_j) = p(X_1 \mid P_j) \cdot ... \cdot p(X_m \mid P_j). \tag{10}$$

Założenie to zazwyczaj nie jest spełnione, dlatego algorytm nazywany jest „naiwnym".

W eksperymencie modele dyskryminacyjne budowano zarówno dla zbiorów danych przed dyskretyzacją, jak i po dyskretyzacji.

3.2. Wykorzystane zbiory danych oraz opis eksperymentu

Zbiory danych dobrano w ten sposób, aby różniły się pod względem liczebności, liczby zmiennych i liczby klas. Wszystkie zbiory dostępne są w odpowiednich pakietach programu **R**.

[1] Decyzja o przydzieleniu obserwacji $\boldsymbol{x}_i$ do klasy P_j podejmowana jest już po zaobserwowaniu $\boldsymbol{x}_i$.

Tabela 1. Charakterystyka zbiorów danych

Zbiór danych	Liczba obserwacji	Liczba zmiennych	Liczba klas
bupa	345	7	2
iris	150	5	3
diabetes	768	8	2
ionosphere	351	33	2
sonar	208	61	2
Satellite	6435	37	6
vehicle	846	19	4

Źródło: opracowanie własne.

Przed przystąpieniem do dyskretyzacji zmiennych dla każdego ze zbiorów zbudowano model dyskryminacyjny dla zbioru uczącego, a następnie wyznaczono błąd klasyfikacji na zbiorze testowym stanowiącym $1/3$ wszystkich obserwacji badanego zbioru.

Kolejna część eksperymentu polegała na badaniu błędu klasyfikacji dla zbioru testowego w zależności od wybranej metody dyskretyzacji[2].

Algorytm dyskretyzacji zmiennej ciągłej składa się z czterech etapów:

1. Uporządkowanie wartości zmiennej.
2. Wyznaczenie odpowiednich punktów podziału lub dwóch sąsiadujących przedziałów.
3. Podział lub połączenie przedziałów (zależy od zastosowanej metody dyskretyzacji).

Powtarzanie kroków 2. i 3. aż do osiągnięcia odpowiedniego dla danej metody kryterium stopu.

Trzy z czterech zastosowanych metod dyskretyzacji wymagają od użytkownika ustalenia wartości pewnych parametrów, tj. liczby przedziałów, na jakie podzielone zostaną wartości danej zmiennej (metody bezkontekstowe), czy poziomu istotności α (metoda kontekstowa ChiMerge). Dyskretyzując zbiory danych metodami kontekstowymi, liczbę przedziałów dla poszczególnych zmiennych ustalono na podstawie formuły Sturgesa, potem Scotta i Freedmana-Diaconisa. W metodzie Chi-Merge dyskretyzacji dokonano w zależności od poziomu istotności α równego 0,01, 0,05, 0,1, oraz 0,5. Następnie każdy zdyskretyzowany zbiór podzielono na część uczącą i testową w stosunku 2 : 1. Dla zbioru uczącego wyznaczono model dyskryminacyjny, wykorzystując naiwny klasyfikator bayesowski, i dla niego obliczono błąd klasyfikacji liczony metodą sprawdzania krzyżowego z podziałem zbioru uczącego na 10 części. Do dalszej analizy, dla każdej z metod dyskretyzacji, wybrano układ parametrów minimalizujący powyższy błąd. Zestawienie wybranych parametrów przedstawia tab. 2.

[2] Losowanie tego samego zestawu obserwacji w zbiorach testowych zapewnia funkcja `sed.seed()`.

Tabela 2. Wybrane parametry dla metod dyskretyzacji

Zbiór danych	Liczba przedziałów		α
	`disc.ew`	`disc.ef`	`chiMerge`
bupa	F. Sturgesa	F. Freedmana-Diaconisa	0,5
diabetes	F. Sturgesa	F. Scotta	0,01
iris	F. Freedmana-Diaconisa	F. Freedmana-Diaconisa	0,01
ionosphere	F. Sturgesa	F. Freedmana-Diaconisa	0,01
sonar	F. Scotta	F. Scotta	0,1
Satellite	F. Sturgesa	F. Scotta	0,01
vehicle	F. Scotta	F. Scotta	0,1

Źródło: opracowanie własne.

W następnym kroku wyznaczono błąd klasyfikacji dla zbioru testowego.

Obliczenia wykonano w programie **R** na podstawie autorskich procedur oraz funkcji z pakietów `dprep`, `e1071`, `grDevices`, `infotheo` oraz `car`.

3.3. Wyniki eksperymentu

W wyniku zastosowania schematu postępowania omówionego w punkcie 3.2 otrzymano błędy klasyfikacji dla zbiorów testowych, które zestawiono w tab. 3.

Tabela 3. Błędy klasyfikacji (w %)

Zbiór danych	Bez dyskret.	`disc.ew`	`disc.ef`	`disc.mentr`	`chiMerge`
bupa	43,48	41,74	**26,96**	33,91	31,70
diabetes	24,61	25,00	24,61	**22,66**	**22,66**
iris	4	4	4	6	**2**
ionosphere	20,51	18,80	17,95	17,09	**16,24**
sonar	30,43	31,88	24,64	**21,74**	**21,74**
Satellite	21,17	**20,89**	23,50	20,98	21,17
vehicle	54,96	53,90	53,19	**44,68**	45,04

Źródło: opracowanie własne. Czcionką pogrubioną wyróżniono najmniejsze błędy klasyfikacji.

4. Podsumowanie

Porównując otrzymane wyniki można stwierdzić, że lepszą jakością, rozumianą jako wielkość błędu klasyfikacji na zbiorze testowym, charakteryzują się modele dyskryminacyjne otrzymane dla zbiorów zdyskretyzowanych. W przypadku pięciu zbiorów danych mniejsze błędy klasyfikacji uzyskano dla metod kontekstowych. Natomiast jakość modeli zbudowanych dla zbiorów danych zdyskretyzowanych za pomocą metod bezkontekstowych jest na ogół gorsza. Wyniki badań empirycznych pozytywnie weryfikują postawione hipotezy badawcze.

Choć lepsze wyniki, w aspekcie wielkości błędu klasyfikacji, uzyskano stosując kontekstowe metody dyskretyzacji, trudno jednoznacznie określić, która z metod jest najlepsza, a która najgorsza. Kolejne badania służyć będą ocenie przydatności konkretnej metody kontekstowej.

Literatura

Bock H.H., Diday E. (red.), *Analysis of Symbolic Data*. Explo*ratory Methods for Extracting Statistical Information from Complex Data*, Springer, Berlin 2000.

Catlett J., *On Changing Continuous Attributes into Ordered Discrete Attributes*, [w:] Y. Kodratoff (red.), *Proceedings of the European Working Session on Learning*, Springer, Berlin 1991, s. 164-178.

Dougherty J., Kohavi R., Sahami M., *Supervised and Unsupervised Discretization of Continuous Features*, Proceedings of the 12th International Conference on Machine Learning, Morgan Kaufmann, San Francisco 1995, s. 194-202.

Fayyad U.M., Irani K.B., *Multi-interval Discretization of Continuous – Valued Attributes for Classifiaction Learning*, Proceedings of the 13th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, San Francisco 1993, s. 1022-1027.

Freedman D., Diaconis P., *On histogram as a density estimator: L_2 theory*, „Probability Theory and Related Fields" 1981, vol. 57, no. 4, s. 453-476.

Gatnar E., *Problemy dyskretyzacji zmiennych*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 874, Wrocław 2000, s.190-198.

Hsu Ch.-N., Huang H.-J.,Wong T.-T., *Why Discretization Works for Naive Bayesian Classifiers*, Proceedings of the 17th International Conference on Machine Learning, Stanford 2000, s. 399-406.

Kerber R., *ChiMerge: Discretization of Numerical Attributes*, Proceedings of the 10th National Conference on Artificial Intelligence, MIT Press, San Jose 1992, s. 123-128.

Scott D. W., *On optimal and data-based histograms*, „Biometrika" 1979, vol. 66, no. 3, s. 605-610.

Sturges H., *The choice of a class-interval*, „Journal of the American Statistical Association" 1926, vol. 21, no. 153, s. 65-66.

Walesiak M., Gatnar E. (red.), *Statystyczna analiza danych z wykorzystaniem programu R*, Wyd. Naukowe PWN, Warszawa 2009, s. 193-194.

Yang Y., Webb G.I., *Proportional k-interval Discretization for Naive-Bayes Classifiers*, Proceedings of the 12th European Conference on Machine Learning, Springer, Berlin 2001, s. 564-575.

THE INFLUENCE OF DISCRETIZATION METOD ON CLASSIFICATION QUALITY

Summary: The aim of this article is to compare classification errors of classification models for data sets before and after discretization. The naive-Bayes classifiers as a supervised `classification method was used`. It was trained on a data before discretization and on a data preprocessed by discretization methods. The unsupervised (discretization using intervals of equal width, discretization using intervals of equal frequencies) and supervised (discretization using the Chi-Merge method, discretization using the minimum entropy criterion) discretization methods are used. In the empirical part, procedures from `dprep`, `e1071`, `grDevices`, `infotheo` and `car` packages for **R** software are used.