AUTOMATYKA • 2010 • Tom 14 • Zeszyt 2

Wiesław Wajs*

Klasyfikacja przypadków medycznych metodą statystyczną

1. Wstęp

W praktyce medycznej występuje problem klasyfikacji danych. Rozważamy zagadnienie, gdy dany jest jeden przypadek chorobowy opisany dwoma parametrami: masą urodzeniową *mu* oraz wiekiem płodowym *wp*. Problem klasyfikacji polega na określeniu, czy dany przypadek chorobowy należy do zbioru etykietowanego „1", czy należy do zbioru etykietowanego „0".

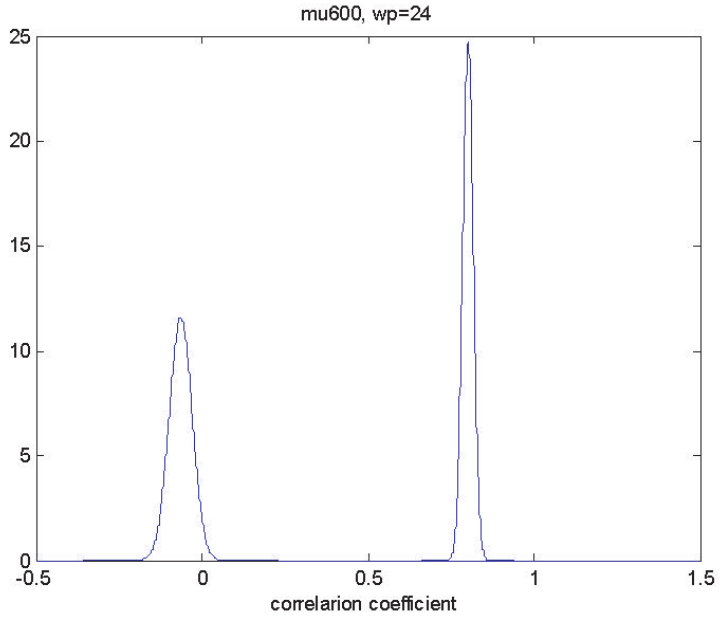
Proponowana w pracy metoda umożliwia obliczenie prawdopodobieństwa, z jakim rozpatrywany przypadek chorobowy należy do zbioru etykietowanego „1", oraz prawdopodobieństwa, z jakim ten przypadek należy do zbioru etykietowanego „0". W naukach medycznych problem klasyfikacji jest szczególnie ważny. Zakwalifikowanie danego przypadku chorobowego do zbioru rozpoznanych przypadków na znaczenie praktyczne. Znane są liczne metody pozwalające zakwalifikować rozpatrywany przypadek do znanego zbioru danych. Ze znanych metod klasyfikacji należy wymienić metodę *Support Vector Machine* [3, 4], metodę sztucznych sieci neuronowych [8] , a także metodę sztucznych sieci immunologicznych [6, 7]. Metody te nie określają prawdopodobieństwa, z jakim dany przypadek należy do zbioru danych. Proponowana metoda wykorzystuje metodę korelacji dwóch parametrów tego samego przypadku chorobowego. Obliczone współczynniki korelacji tworzą rozkład gęstości prawdopodobieństwa. Badany jest wpływ każdej pary parametrów na rozkład gęstości prawdopodobieństwa. Badanie wpływu każdej pary parametrów na rozkład gęstości prawdopodobieństwa polega na usuwaniu ze zbioru danych po jednej parze parametrów, za każdym razem innej pary parametrów, w celu uzyskania rozkładu za pomocą algorytmu trenowania. Obliczony w taki sposób rozkład gęstości prawdopodobieństwa jest wykorzystywany przez algorytm trenujący.

Algorytm testu metody oblicza wpływ, jaki na rozkład gęstości prawdopodobieństwa ma jeden przypadek określony dwoma parametrami oznaczonymi *mu* oraz *wp*. Testowanie polega na zastąpieniu w zbiorze trenującym każdej pary parametrów przez testową parę

* Katedra Automatyki, Akademia Górniczo-Hutnicza w Krakowie

parametrów. Rozkład gęstości prawdopodobieństwa obliczny jest dwukrotnie. Pierwszy raz za pomocą algorytmu trenowania, a drugi raz za pomocą algorytmu testowania.

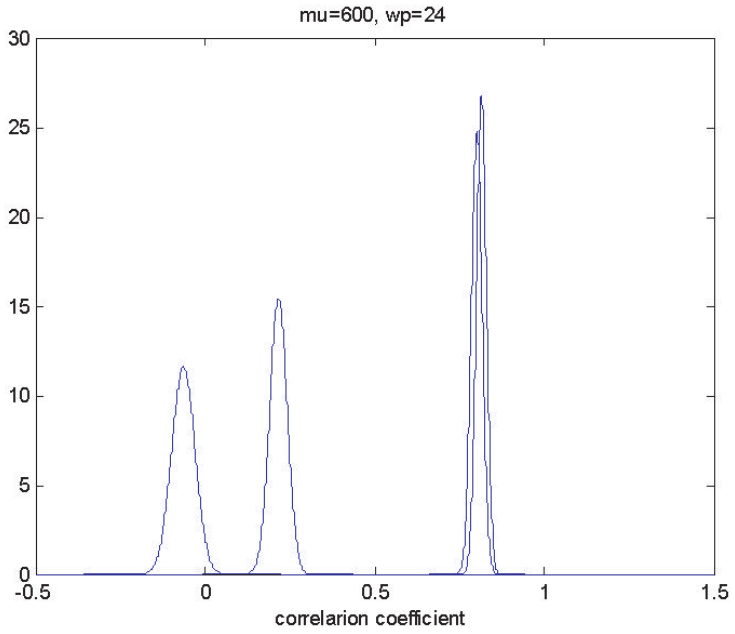
Obliczone rozkłady gęstości prawdopodobieństwa za pomocą algorytmu trenowania można porównać z rozkładem obliczonym za pomocą algorytmu testowania.



Rys. 1. Rozkład gęstości prawdopodobieństwa uzyskany za pomocą algorytmu trenowania dla parametrów *mu* = 600 i *wp* = 24, obliczony dla zbioru etykietowanego „0” (ukazany na rysunku na lewo) oraz dla zbioru etykietowanego „1” w prawej części rysunku

Na rysunku 1 przedstawiono dwa wykresy rozkładów gęstości prawdopodobieństwa współczynników korelacji. Wykres położony na lewo obrazuje rozkład gęstości prawdopodobieństwa obliczony za pomocą algorytmu trenowania dla danych opisujących przypadki, dla których nie wystąpiło schorzenie. Na prawo od tego wykresu znajduje się wykres gęstości prawdopodobieństwa obliczony za pomocą algorytmu trenowania dla przypadków, u których wystąpiło schorzenie. Te przypadki są etykietowane „1”.

Na rysunku 2 znajdują się cztery wykresy. Dwa wykresy przedstawiono już na rysunku 1. Wykresy położone w lewej części rysunku 2 mają znikomą wspólną powierzchnię, a dwa inne wykresy, położone na prawo mają znaczną wspólną powierzchnię. Nowe wykresy uwidocznione na rysunku 2 uzyskano za pomocą algorytmu testowania. Obserwując wykresy gęstości prawdopodobieństwa, można wnioskować, że przypadek testowy określony parametrami *mu* = 600 i *wp* = 24 nie należy do zbioru przypadków etykietowanego „0”, ale należy do zbioru przypadków etykietowanych „1”.



Rys. 2. Rozkład gęstości prawdopodobieństwa uzyskany za pomocą algorytmów trenowania i testowania dla parametrów *mu* = 600 i *wp* = 24

Wartość średnia dla tego wykresu $mean_{tr1}$ = 0,8007. Wykres gęstości prawdopodobieństwa obliczony algorytmem testowym znajduje się jako pierwszy po prawej stronie. Wartość średnia dla tego wykresu $mean_{te1}$ = 0,8141. Dla tych wykresów część wspólna jest znacza i wynosi około 0,7. Prawdopodobieństwo z jakim dany przypadek należy do zbioru etykietowanego „1", można obliczyć operacją całkowania. Granice odpowiednich całek wyznaczają dwa punkty przecięcia odpowiednich wykresów gęstości prawdopodobieństwa.

2. *Bronchopulmonary dysplasia* (bpd)

Podstawowym celem klasyfikacji jest określenie prawdopodobieństwa wystąpienia bpd w drugiej dobie życia noworodka na podstawie danych dwóch parametrów: *mu* oraz *wp*. Należy zauważyć, że samo schorzenie wystąpi ewentualnie dopiero po okresie 3–4 tygodni od terminu klasyfikacji. *Bronchopulmonary dysplasia* jest chronicznym schorzeniem płuc występującym u noworodków poddanym wentylacją mechaniczną z użyciem respiratora. Schorzenie to występuje szczególnie wsród noworodków z małą masą urodzeniową, mniejszą niż 1500 g, oraz z wiekiem płodowym mniejszym niż 34 tygodnie życia. U dzieci tych, które są poddane mechanicznemu wspomaganiu oddychania, występuje syndrom znany w litraturze medycznej pod nazwą *Respiratory Distress Syndrome*. Sztuczna wentylacja z użyciem respiratora umożliwia przeżycie noworodka, ale równocześnie może

prowadzić do uszkadzania delikatnych pęcherzyków płucnych, zwłaszcza w przypadku gdy występuje przez dłuższy okres czasu.

Gdy symptomy *Respiratory Distress Syndrome* występują trwale można rozpatrywać prawdopodobieństwo wystąpienia schorzenia *Bronchopulmonary Displasia*. Ważnymi czynnikami schorzenia są: wcześniactwo, infekcje, mechaniczna wentylacja. W pracy splatają się dwa problemy. Jest to problem klasyfikacji i problem predykcji. Problem predykcji występuje dlatego, ponieważ klasyfikację rozpatrujemy w drugiej dobie życia, a okres predykcji obejmuje 3–4 tygodni.

3. Metoda klasyfikacji

Wproponowanej metodzie klasyfikacji zakłada się normalny rozkład gęstości prawdopodobieństwa dla współczynników korelacji. Wprzypadku gdy wykres gęstości prawdopodobieństwa obliczony algorytmem testowym znajduje się na prawo od wykresu gęstości prawdopodobieństwa obliczonego za pomocą algorytmu trenowania, można napisać warunek klasyfikacji w postaci nierówności (4).

Dla zbioru etykietowanego „1” dane są dwa rozkłady gęstości prawdopodobieństwa, jeden obliczony algorytmem trenowania, adrugi obliczony algorytmem testowania. Wartości średnie i odchylenia standardowe oznaczymy odpowiednio: $mean_{tr1}$ i $mean_{te1}$ oraz std_{tr1} i std_{te1}. Punkty przecięcia tych dwóch wykresów gęstości prawdopodobieństwa oznaczymy $x1_1$ oraz $x2_1$.

Dla zbioru etykietowanego „0” dane są dwa rozkłady gęstości prawdopodobieństwa, jeden uzyskany za pomocą algorytmu trenowania, a drugi za pomocą algorytmu testowania. Podobnie dla zbioru etykietowanego „1”, dane są dwa rozkłady gęstości prawdopodobieństwa. Wartości średnie i odchylenia standardowe oznaczymy odpowiednio: $mean_{tr0}$ i $mean_{te0}$ oraz std_{tr0} i std_{te0}. Punkty przecięcia tych dwóch wykresów oznaczymy $x1_0$, $x2_0$.

3.1. Granice całkowania dla rozkładów gęstości prawdopodobieństwa

Postawiony problem polega na obliczeniu prawdopodobieństwa, z jakim przypadek testowy opisany parametrami mu oraz wp należy do zbioru etykietowanego „1”. Obliczenie miejsc przecięcia wykresów gęstości prawdopodobieństwa uzyskanych algorytmami trenowania i testowania pozwala określić granice całkowania konieczne do obliczenia prawdopodobieństwa.

Dwa wykresy gęstości prawdopodobieństwa mają dwa, $x1_1$ oraz $x2_1$, punkty przecięcia przy założeniu, że mają różne wartości średnie i różne wartości odchylenia standardowego. Oznaczając parametry równania kwadratowego przez a, b, c, wartości te można obliczyć ze wzorów na pierwiastki $x1$, $x2$ równania kwadratowego w następujący sposób.

Porównamy funkcje

$$\frac{1}{std_{te1}\sqrt{(2\pi)}}\exp\left(\frac{-(x-mean_{te1})^2}{2std_{te1}^2}\right)=\frac{1}{std_{tr1}\sqrt{(2\pi)}}\exp\left(\frac{-(x-mean_{tr1})^2}{2std_{tr1}^2}\right).$$

Po zlogarytmowaniu otrzymamy

$$\ln\frac{1}{std_{te1}\sqrt{(2\pi)}}+(\frac{-(x-mean_{te1})^2}{2std_{te1}^2})=\ln\frac{1}{std_{tr1}\sqrt{(2\pi)}}+(\frac{-(x-mean_{tr1})^2}{2std_{tr1}^2}).$$

Stąd

$$-(x^2-2mean_{tr1}x+mean_{tr1}^2)std_{te1}^2+$$

$$(x^2-2mean_{te1}x+mean_{te1}^2)std_{tr1}^2-$$

$$2std_{tr1}^2std_{te1}^2(\ln\frac{1}{std_{te1}\sqrt{(2\pi)}}-\ln\frac{1}{std_{tr1}\sqrt{(2\pi)}})=0.$$

Współczynniki równania kwadratowego a, b i c dane są wzorami

$$a=std_{tr}^2-std_{te}^2 \tag{1}$$

$$b=2(mean_{tr}std_{te}^2-mean_{te}std_{tr}^2) \tag{2}$$

$$c=2std_{tr}^2std_{te}^2(\ln\frac{1}{std_{tr}\sqrt{(2\pi)}}-\ln\frac{1}{std_{te}\sqrt{(2\pi)}})-mean_{te}^2std_{tr}^2+mean_{tr}^2std_{te}^2 \tag{3}$$

Załóżmy bez utaty ogólności, że wykresy uzyskane za pomocą algorytmów testowych są położone na prawo od wykresów uzyskanych za pomocą algorytmów trenowania. Kryterium klasyfikacji można sformułować w postaci: jeżeli spełniona jest relacja (4) dla arbitralnie dobranej wartości dodatniej Δ, to przypadek testowy spełnia warunek przynależności do zbioru przypadków etykietowanych „1”

$$\frac{1}{std_{te1}\sqrt{(2\pi)}}\int_{-1}^{x1_1}\exp(\frac{-(x-mean_{te1})^2}{2std_{te1}^2})dx+$$

$$\frac{1}{std_{tr1}\sqrt{(2\pi)}}\int_{x1_1}^{x2_1}\exp(\frac{-(x-mean_{tr1})^2}{2std_{tr1}^2})dx+$$

$$\frac{1}{std_{te1}\sqrt{(2\pi)}}\int_{x2_1}^{1}\exp(\frac{-(x-mean_{te1})^2}{2std_{te1}^2})dx-$$

$$\frac{1}{std_{te0}\sqrt{(2\pi)}}\int_{-1}^{x1_0}\exp(\frac{-(x-mean_{te0})^2}{2std_{te0}^2})dx+$$

$$\frac{1}{std_{tr0}\sqrt{(2\pi)}}\int_{x1_0}^{x2_0}\exp(\frac{-(x-mean_{tr0})^2}{2std_{tr0}^2})dx+$$

$$\frac{1}{std_{te0}\sqrt{(2\pi)}}\int_{x2_0}^{1}\exp(\frac{-(x-mean_{te0})^2}{2std_{te0}^2})dx\geq\Delta.$$

Warunek (4) można zapisać w zależności od wzajemnego położenia wykresów gęstości prawdopodobieństwa współczynników korelacji. Wzajemne położenie wykresów gęstości prawdopodobieństwa określają obliczone wartości średnie. Cztery układy wzajemnego położenia wykresów gęstości prawdopodobieństwa zależą od wartości obliczonych średnich dla trenowania i testowania

$$mean_{te0} > mean_{tr0} \text{ and } mean_{te1} > mean_{tr1} \tag{5}$$

$$mean_{te0} < mean_{tr0} \text{ and } mean_{te1} < mean_{tr1} \tag{6}$$

$$mean_{te0} > mean_{tr0} \text{ and } mean_{te1} < mean_{tr1} \tag{7}$$

$$mean_{te0} < mean_{tr0} \text{ and } mean_{te1} > mean_{tr1} \tag{8}$$

Położenie wykresów gęstości prawdopodobieństwa współczynników korelacji ma wpływ na określenie kolejności operacji całkowania dla postaci warunku klasyfikacji (4) i dla odpowiednich wykresów gęstości prawdopodobieństwa. Na przykład dla warunku (6) relację (4) można napisać w postaci

$$\frac{1}{std_{tr1}\sqrt{(2\pi)}}\int_{-1}^{x1_1}\exp(\frac{-(x-mean_{tr1})^2}{2std_{tr1}^2})dx +$$

$$\frac{1}{std_{te1}\sqrt{(2\pi)}}\int_{x1_1}^{x2_1}\exp(\frac{-(x-mean_{te1})^2}{2std_{te1}^2})dx +$$

$$\frac{1}{std_{tr1}\sqrt{(2\pi)}}\int_{x2_1}^{1}\exp(\frac{-(x-mean_{tr1})^2}{2std_{tr1}^2})dx -$$

$$\frac{1}{std_{tr0}\sqrt{(2\pi)}}\int_{-1}^{x1_0}\exp(\frac{-(x-mean_{tr0})^2}{2std_{tr0}^2})dx +$$

$$\frac{1}{std_{te0}\sqrt{(2\pi)}}\int_{x1_0}^{x2_0}\exp(\frac{-(x-mean_{te0})^2}{2std_{te0}^2})dx +$$

$$\frac{1}{std_{tr0}\sqrt{(2\pi)}}\int_{x2_0}^{1}\exp(\frac{-(x-mean_{tr0})^2}{2std_{tr0}^2})dx \geq \Delta.$$

3.2. Algorytm trenowania

Algorytm trenowania jest użyty do obliczenia współczynników korelacji. Zbiór współczynników korelacji tworzy rozkład gęstości prawdopodobieństwa. Obliczamy dwa rozkłady gęstości prawdopodobieństwa. Pierwszy rozkład gęstości prawdopodobieństwa obliczamy dla danych, co do których mamy pewność, że występuje dla nich schorzenie bpd, dane

te są etykietowane „1". Drugi rozkład gęstości prawdopodobieństwa obliczamy dla danych etykietowanych „0". Dla tych danych mamy pewność, że nie opisują one schorzenia bpd.

Zakłada się, że wartość średnia rozkładu gęstości prawdopodobieństwa otrzymania dla zbioru danych etykietowanych „1" jest różna od wartości średniej rozkładu gęstości prawdopodobieństwa otrzymana dla zbioru danych etykietowanych „0". Zakładamy arbitralnie, że im większa jest różnica pomiędzy wartościami średnimi dla zbiorów, oraz im mniejsze są wartości odchyleń standardowych, tym metoda klasyfikacji jest bardziej skuteczna.

Równanie współczynnika korelacji wykorzystywane do obliczenia odpowiednich rozkładów gęstości prawdopodobieństwa dane jest wzorem (10).

$$\rho_{mu,wp} = \frac{\mathrm{cov}(mu, wp)}{std_{mu}\, std_{wp}} \tag{10}$$

$$-1 <= \rho_{mu,wp} <= 1$$

$$\mathrm{cov}(mu, wp) = \frac{1}{n}\sum_{i=1}^{i=n}(mu_i - mean_{mu})(wp_i - mean_{wp}).$$

3.3. Algorytm testowania

Algorytm testowy umożliwia obliczenie prawdopodobieństwa, z jakim dany przypadek testowy należy do zbioru etykietowanego „1" oraz do zbioru etykietowanego „0". Algorytm testowy oblicza dwa rozkłady gęstości prawdopodobieństwa. Do uzyskania rozkładu gęstości prawdopodobieństwa współczynników korelacji stosuje się równanie (10). Rozkład gęstości prawdopodobieństwa obliczamy algorytmem testowym, zastępując każdą parę parametrów w zbiorze danych zródłowych wybranym przypadkiem testowym *mu*, *wp*.

4. Analiza przypadku *mu* = 600 *wp* = 24

Zebrano dane opisujące schorzenie bpd w postaci zbioru par parametrów mu oraz wp. Zbiór danych źródłowych obejmuje 52 pary danych dla algorytmu trenowania i kilka par danych dla algorytmu testowego. Dane źródłowe zamieszczono w tabeli 1. W tabeli 1 zamieszczono 26 par parametrów, które opisują przypadki, co do których jest pewność wystąpienia schorzenia bpd etykietowanego „1". Ponadto, w tabeli 1 zamieszczono 26 par parametrów dla zbioru etykietowanego „0". Aby uzyskać odpowiednie rozkłady prawdopodobieństwa, stosujemy algorytm trenowania i algorytm testowania.

Algorytm trenowania zastosujemy 26 razy dla zbioru etykietowanego „1" i 26 razy dla zbioru etykietowanego „0". Za każdym razem zastosowania algorytmu trenowania usuwamy inną parę danych ze zbioru etykietowanego „0" obliczamy 26 współczynników korelacji ρ z równania (10). Podobnie dla zbioru etykietowanego „1" obliczymy 26 wartości dla współczynnika korelacji.

Tabela 1

Dane źródłowe

Nr	*mu* Label 1 bpd	*wp* Label 1 bpd	*mu* Label 0 no bpd	*wp* Label 0 no bpd
1	890	25	1360	29
2	700	24	1400	31
3	1100	28	880	28
4	760	28	985	32
5	1200	29	1100	28
6	700	25	1100	28
7	960	26	1300	29
8	760	25	900	28
9	860	27	1100	30
10	600	24	1400	30
11	860	28	1100	28
12	1300	29	1000	30
13	1400	31	880	30
14	940	28	985	32
15	800	28	1100	32
16	1000	27	1100	27
17	600	25	1300	27
18	950	28	900	29
19	1095	28	1100	32
20	800	25	1400	28
21	760	25	1100	26
22	770	24	1200	28
23	1500	30	900	28
24	1200	30	1250	31
25	650	28	1250	30
26	720	25	930	32

Wyniki tych obliczeń zamieszczono w tabeli 2. Wartości zawarte w wierszu numer 1 w tabeli 2 obliczono ze wzoru (10), gdy w tabeli 1 usunięto wiersz numer 1. Odpowiednio usunięto parametry *mu* = 890 i *wp* = 25 ze zbioru etykietowanego „1”, oraz parametry *mu* = 1360 i *wp* = 29 ze zbioru etykietowanego „0”.

W tabeli 3 zamieszczono wyniki działania algorytmu testowego. Algorytm testowy wymaga zastąpienia każdej pary parametrów zamieszczonych w tabeli 1 przez parametry przypadku testowego *mu* = 600 i *wp* = 24. Dane zamieszczone w tabeli 3 tworzą rozkład gęstości prawdopodobieństwa opisujący zmianę, jaką w zbiorze danych źródłowych powoduje zastąpienie każdej z 26 par parametrów przez parę parametrów testowych: *mu* = 600, *wp* = 24. Wartości zamieszczone w wierszu numer 1 w tabeli 3 obliczono ze wzoru (10), gdy w tabeli 1 zastąpiono w wierszu numer 1 odpowiednio parametry *mu* = 890 i *wp* = 25 parametrami *mu* = 600 i *wp* = 24 dla zbioru etykietowanego „1”, a parametry *mu* = 1360 i *wp* = 29 zastąpiono parametrami *mu* = 600 i *wp* = 24 dla zbioru etykietowanego „0”.

Podobnie, w wierszu numer 2 zastąpiono parametry *mu* = 700 i *wp* = 24 parametrami *mu* = 600 i *wp* = 24 dla zbioru etykietowanego „1”, tu parametry *mu* = 1400 i *wp* = 31 zastąpiono parametrami *mu* = 600 i *wp* = 24 dla zbioru etykietowanego „0”.

Tabela 2

Wartości współczynnika korelacji $\rho_{mu,wp}$ obliczony algorytmem trenowania dla zbioru etykietowanego „1” i dla zbioru etykietowanego „0”

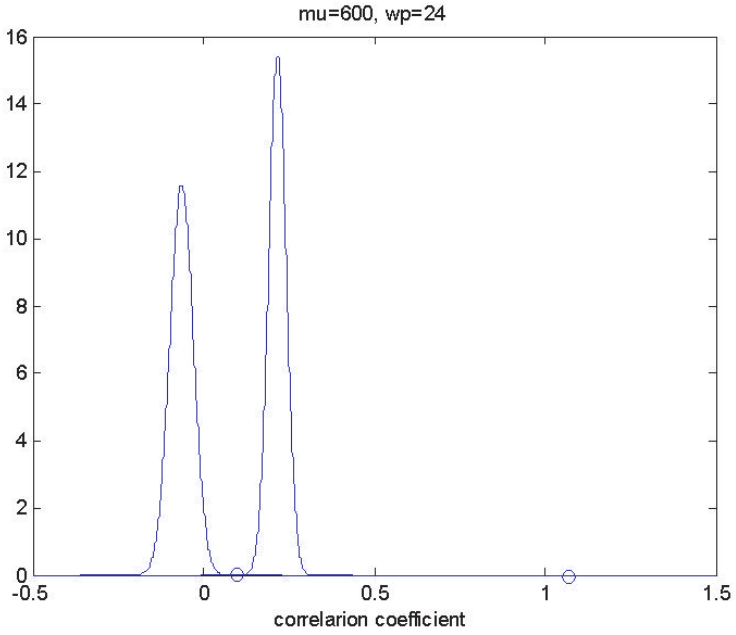
Nr	ρ *mu* , *wp* Label 1	ρ *mu* , *wp* Label 0
1	0,8113	–0,0551
2	0,7948	–0,1375
3	0,7985	–0,1126
4	0,8269	–0,0183
5	0,7910	–0,0678
6	0,7936	–0,0678
7	0,8078	–0,0569
8	0,7971	–0,1080
9	0,8021	–0,0629
10	0,7844	–0,0948
11	0,8116	–0,0678
12	0,7926	–0,0545
13	0,7622	–0,0452
14	0,8036	–0,0183
15	0,8200	–0,0612
16	0,8021	–0,0718
17	0,7926	–0,0064
18	0,8029	–0,0767
19	0,7983	–0,0612
20	0,8005	–0,0136
21	0,7971	–0,0772
22	0,8054	–0,0497
23	0,7983	–0,1080
24	0,7877	–0,0968
25	0,8511	–0,0770
26	0,7946	0,0026

Tabela 3

Wartości współczynnika korelacji $\rho_{mu,wp}$ obliczony algorytmem testowania dla zbioru etykietowanego „1” i dla zbioru etykietowanego „0”

Nr	ρ *mu* , *wp* Label 1	ρ *mu* , *wp* Label 0
1	0,8240	0,2201
2	0,8097	0,1625
3	0,8114	0,1891
4	0,8385	0,2517
5	0,8046	0,2098
6	0,8083	0,2098
7	0,8204	0,2158
8	0,8112	0,1906
9	0,8152	0,2086
10	0,8008	0,1923
11	0,8240	0,2098
12	0,8062	0,2182
13	0,7787	0,2332
14	0,8164	0,2517
15	0,8320	0,2159
16	0,8150	0,2136
17	0,8080	0,2632
18	0,8156	0,2095
19	0,8112	0,2159
20	0,8142	0,2570
21	0,8112	0,2199
22	0,8191	0,2226
23	0,7991	0,1906
24	0,8015	0,1851
25	0,8615	0,1982
26	0,8091	0,2704

Punkty przecięcia wykresów gęstości prawdopodobieństwa (rys. 3, 4) oznaczone na rysunkach znakiem „o” dla danych etykietowanych „0” to $x1_0 = 0{,}0951$ oraz $x2_0 = 1{,}0688$. Punkty przecięcia oznaczone znakiem „*” dla danych etykietowanych „1” to $x1_1 = 0{,}8062$ oraz $x2_1 = 0{,}9802$. Punkty te umożliwiają obliczenie prawdopodobieństwa, z jakim przypadek opisany parametrami *mu* = 600 i *wp* = 24 należy do zbioru danych etykietowanych „1” oraz prawdopodobieństwa, z jakim dany przypadek należy do zbioru danych etykietowanych „0”.



Rys. 3. Rozkład gęstości prawdopodobieństwa uzyskany za pomocą algorytmów trenowania i testowania dla parametrów *mu* = 600 i *w*p = 24, obliczony dla zbioru etykietowanego „0", znakiem „o" oznaczono dwa punkty przecięcia wykresów gęstości prawdopodobieństwa

Część wspólną dla obu wykresów po prawej stronie rysunku 4 można obliczyć za pomocą trzech całek, poprzez całkowanie funkcji gęstości prawdopodobieństwa. Dodając wartości: 0,29798 + 0,361655 + 0,119188·10^{-28} otrzymamy prawdopodobieństwo, w jakim dany przypadek *mu* = 600 i *wp* = 24 należy do zbioru etykietowanego „1":

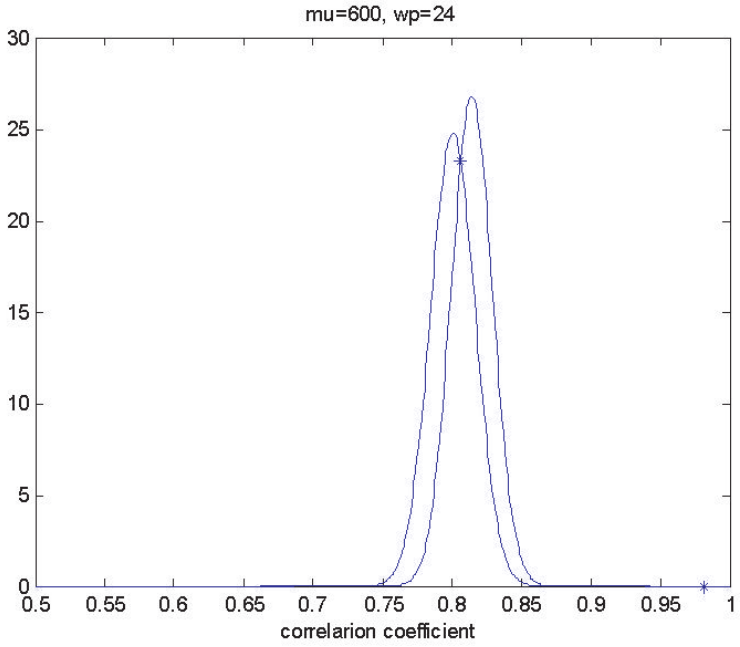
$$\frac{1}{0,0149\sqrt{(2\pi)}}\int_{-1}^{0,8062}\exp(\frac{-(x-0,8141)^2}{2\cdot 0,0149^2})dx = 0,29798 \tag{11}$$

$$\frac{1}{0,0161\sqrt{(2\pi)}}\int_{-0,8062}^{0,9802}\exp(\frac{-(x-0,8007)^2}{2\cdot 0,0161^2})dx = 0,361655 \tag{12}$$

$$\frac{1}{0,0149\sqrt{(2\pi)}}\int_{0,9802}^{1}\exp(\frac{-(x-0,8141)^2}{2\cdot 0,0149^2})dx = 0,119188\cdot 10^{-28} \tag{13}$$

Dwa punkty przecięcia wykresów dla zbiorów etykietowanych „1" to $x1_1$ = 0,8062 i $x2_1$ = 0,9802. Wartość średnia dla gęstości prawdopodobieństwa obliczona algorytmem trenowania dla zbioru etykietowanego „1" to $mean_{tr1}$ = 0,8007. Wartość obliczona algoryt-

mem testowania $mean_{te1}$ = 0,8141. Wartość std obliczona algorytmem testowym dla zbioru etykietowanego „1” to std_{te1} = 0,0149. Wartość std obliczona algorytmem trenowania dla zbioru etykietowanego „1” to std_{tr1} = 0,0161.



Rys. 4. Rozkład gęstości prawdopodobieństwa uzyskany za pomocą algorytmów trenowania i testowania dla parametrów *mu* = 600 i *wp* = 24, obliczony dla zbioru etykietowanego „1”, znakiem „*” oznaczono dwa punkty przecięcia wykresów gęstości prawdopodobieństwa

5. Podsumowanie

Za pomocą prezentowanej metody klasyfikacji można rozpatrywać cztery ogólne przypadki.

Pierwszy przypadek dotyczy sytuacji, gdy wykresy gęstości prawdopodobieństwa dla przypadków etykietowanych „1” uzyskane algorytmem trenowania i testowania mają znaczącą część wspólną, a odpowiednie wykresy dla zbiorów etykietowanych „0” mają małą część wspólną. Dla tego przypadku można stwierdzić z określonym prawdopodobieństwem, że ten przypadek należy do zbioru przypadków etykietowanych „1”.

Drugi przypadek ogólny obejmuje sytuacje, gdy wykres gęstości prawdopodobieństwa ma znaczną część wspólną dla zbioru etykietowanego „0” i małą część wspólną dla zbioru etykietowanego „1”. W tym przypadku istnieje duże prawdopodobieństwo, że rozpatrywany przypadek należy do zbioru etykietowanego „0”.

Trzeci przypadek ogólne obejmuje sytuację, gdy wykresy gęstości prawdopodobieństwa mają znaczną część wspólną zbiorów etykietowanych „1", oraz znaczną część wspólną dla zbiorów etykietowanych „0". Wtym przypadku prawdopodobieństwo można obliczyć na podstawie różnicy wartości odpowiednich całek obliczonych z wykorzystaniem funkcji gęstości prawdopodobieństwa.

Czwarty przypadek obejmuje sytuację, gdy wykresy gęstości prawdopodobieństwa mają małą część wspólną dla zbiorów etykietowanych „0" i mają małą część wspólną dla zbiorów etykietowanych „1".

Dla zebranych danych oryginalnych tylko pierwszy przypadek wskazuje jednoznacznie na przynależność rozpatrywanego przypadku testowego do zbioru przypadków etykietowanych „1". Pozostałe trzy przypadki umożliwiają określenie prawdopodobieństwa.

Literatura

[1] Vapnik V., Chervonenkis A., *The necessary and sufficient conditions for con-sistency in the empirical risk minimalization method*. Pattern Recognition and Image Analysis, 1(3), 1991, 283–305.

[2] Mangasarian O.L., *Linear and nonlinear separation of patterns by linear programming*. Operations Research, 13, 1965, 444–452.

[3] Vapnik V., *Statistical Learning Theory*. Wiley, New Jork 1998.

[4] Smola A.J., Barlet P.L., Scholkopf B., Schuurmans D., *Advances in Large Margin Classifiers*. The MTI Press, Cambridge, Massachusetts 2000.

[5] Castro de L.N., von Zuben F.J., *An evoluntary immune system network for data clustering*. In Proc. 6th Brazilian symp. Neural Network. Rio de Janeiro, Brazil 2000, 84–89.

[6] Castro de L.N., von Zuben F.J., *Learning and optimization using the clonal selection principle*. IEEE Tran. Evol. Comput., vol. 6, No. 3, Jan. 2002, 239–251.

[7] Jerne N.K., *Towards a network theory of the immune system*. Annu. Immunol. (Inst Pasteur), vol. 125, No. C, 1974, 373–389.

[8] Demuth H., Beale M., *Neural Network Toolbox: For use with MATLAB: user's Guide*. The Mathworks, 1993.

[9] Bradley A.P., *The use of the area under the ROC curve in the evaluation on machine learning algorithms*. Pattern Recognition, 30(7), 1997, 1145–1159.

[10] Cunha G.S., Mezzacappa-Fihlo F., Ribeiro J.D., *Risk Factors for Bronchopul-monary Dysplasia in very Low Birth Weight Nowborns Treated with Mechanical Venti-lation in the First Week of Life*. Journal of Tropical Perdiatrics, 51(6), 2005, 334–340.

[11] Fawcett T., *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*. Technical Report HPL-2003-4, HP Labs, 2003.

[12] Hanley J.A., McNeil B.J., *The meaning and use of the area under a receiver operating characteristic (ROC) curve*. Radiology, 1982, 29–36.

[13] Hosmer D., Lemeshow S., *Applied Logistic Regresion*. John Wiley and Sons Inc., 1998

[14] Kleinbaum D.G., Klain M., *Logistic Regression – A Self-Learning Text*. Springer-Verlag, New York 2002.

[15] Moody J., Darken C.J., *Fast learning in networks of locally-tuned processing units*. Neural Computation, 1989, 281–294.

[16] Poggio T., Girosi F., *Network for approximation and learning*. Proceedings of the IEEE, 1990, 1481–1497.

[17] Tapia J.L., Agost D., Alegria A., Standen J., Escobar E., Grandi C., Musante G., Zegarra J., Estay A., Ramirez R., *Bronchopulmonary displasia: incidence, risk factors and resource utilization in a population of South-American very low birth weight infants*. J. Pediatric. vol. 82 No. 1, 2006.

[18] Wajs W., Stoch P., Kruczek P., *Bronchopulmonary Dysplasia Prediction using Lo-gistic Regresion*. Proc. of the Sixth International IEEE Conference on Intelligent Systems Design and Applications (ISDA'06). vol. 03, 2006, 98–102.

[19] Fabian J., Farbiaz J., Alvarez D., Martinez C., *Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room*. BioMed Central, Feb. 17, 2005.