INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Wybrane problemy
Tom 11

Pod redakcją
Jerzego HOŁUBCA

Warszawa 2009

INSTYTUT BADAŃ SYSTEMOWYCH
POLSKIEJ AKADEMII NAUK

ANALIZA SYSTEMOWA W FINANSACH I ZARZĄDZANIU

Wybrane problemy
Tom 11

Pod redakcją
Jerzego HOŁUBCA

Warszawa 2009

Wykaz opiniodawców artykułów zamieszczonych w niniejszym tomie:

prof. dr hab. inż. Jerzy HOŁUBIEC
dr inż. Lech KRUŚ
doc. dr hab. inż. Wiesław KRAJEWSKI
doc. dr hab. Jacek MALINOWSKI
dr inż. Edward MICHALEWSKI
prof. dr Adam SKOREK
dr hab. Ryszard SMARZEWSKI
prof. dr hab. inż. Andrzej STRASZAK
dr Dominik ŚLĘZAK
prof. dr hab. inż. Stanisław WALUKIEWICZ
doc. dr hab. Sławomir ZADROŻNY

ISBN 9788389475220

Druk: Zakład Poligraficzny Jerzy Kosiński, Warszawa

ZASTOSOWANIE ZESPOŁU KLASYFIKATORÓW W ZADANIU WSTĘPNEJ SELEKCJI DANYCH

Marcin Gromisz
Studia Doktoranckie IBS PAN

W artykule proponowana jest procedura selekcji (klasyfikacji) wstępnej danych gromadzonych podczas eksperymentów fizycznych. Wysoka wymiarowość przestrzeni reprezentacji i znaczne zróżnicowanie cech danych czynią zadanie konstrukcji odpowiedniego klasyfikatora trudnym do realizacji. W celu rozwiązania tego problemu proponuje się zastosowanie zespołów klasyfikatorów binarnych i oryginalną metodę fuzji ich wyników z użyciem teorii Dempstera-Shafera.

***Słowa kluczowe**: selekcja danych, klasyfikacja, zespół klasyfikatorów, fuzja klasyfikatorów, teoria Dempstera-Shafera*

1. Wprowadzenie

Wiele zjawisk będących przedmiotem zainteresowania fizyki współczesnej cechuje się względną rzadkością występowania w przyrodzie. Aby je zbadać należy dokonać przeglądu dużego obszaru przestrzeni, lub prowadzić obserwacje dostatecznie długo. W obydwu tych sytuacjach masowo gromadzone są dane. Ze względów praktycznych ich ostateczną interpretację musi więc poprzedzać selekcja wstępna, prowadzona tak, by na wczesnym etapie przetwarzania odrzucić jak największą ilość obserwacji nieistotnych dla badacza. Stosowaną w tym celu procedurę diagnostyczną można rozważać jako klasyfikator binarny przypisujący dane (obserwacje) do jednej z dwóch klas decyzyjnych ze zbioru $D = \{akceptacja, dyskwalifikacja\}$.

Szczególnym zagadnieniem rozważanym w pracy są eksperymenty z dziedziny fizyki cząstek elementarnych, w których selekcjonuje się dane napływające z aparatury detekcyjnej, poszukując pewnych ich postaci, świadczących o wystąpieniu określonego rodzaju cząstek. Obecnie większość stosowanych w tym celu procedur, to algorytmy opracowane na podstawie modeli tłumaczących teoretycznie strukturę badanych zjawisk fizycznych. Alternatywą dla nich mogą być algorytmy oparte na analizie obserwowanych danych (ang. *Data driven*), stworzone na podstawie przykładów wyselekcjonowanych poprawnie obiektów. Motywacją takiego podejścia jest przypuszczenie, iż algorytmy o konstrukcji opartej bezpo-

średnio na przetwarzanych danych, pełniej wykorzystują informacje w nich zawarte i mogą być lepsze niż algorytmy wywiedzione a priori z modeli.

Przeszkodą przy konstruowaniu dla omawianych zastosowań klasyfikatorów na podstawie zbiorów przykładów jest charakterystyka selekcjonowanych danych: wielka liczba wymiarów przestrzeni ich reprezentacji ($> 10^3$), niska względna częstość występowania w populacji obiektów poszukiwanych w stosunku do pozostałych ($< 1/10^6$), przy jednoczesnym znacznym zróżnicowaniu ich obserwowanych postaci w obrębie każdej z tych dwóch klas. Przy takich uwarunkowaniach zadanie przygotowania klasyfikatora, uwzględniające reprezentatywny dla całej populacji, odpowiednio liczny zbiór przykładów, może osiągnąć złożoność przekraczającą zdolność obliczeniową powszechnie dostępnych komputerów.

W artykule zaproponowano rozwiązanie tego problemu poprzez pewne przekształcenie oryginalnego zadania i użycie zespołu klasyfikatorów. Każdy z nich jest konstruowany na podstawie jedynie fragmentu zbioru uczącego, jaki byłby niezbędny do budowy klasyfikatora jednolitego. Użycie zespołu klasyfikatorów rodzi problem łącznego uwzględnienia ich odpowiedzi (fuzji) w celu określenia decyzji końcowej o akceptacji albo dyskwalifikacji danych. W proponowanym rozwiązaniu traktuje się to zagadnienie jako zadanie kombinowania przesłanek we wnioskowaniu prowadzonym według zasad *teorii Dempstera-Shafera* (Shafer, 1976).

W punktach 2 i 3 artykułu omówiono kolejno: sposób w jaki przekształcono oryginalne zadanie selekcji oraz fuzję odpowiedzi klasyfikatorów. Punkt 4 przedstawia wyniki eksperymentu obliczeniowego, w którym sprawdzono skuteczność proponowanej metody wstępnej selekcji danych. Rozważania podsumowuje punkt 5.

2. Proponowana procedura selekcji wstępnej

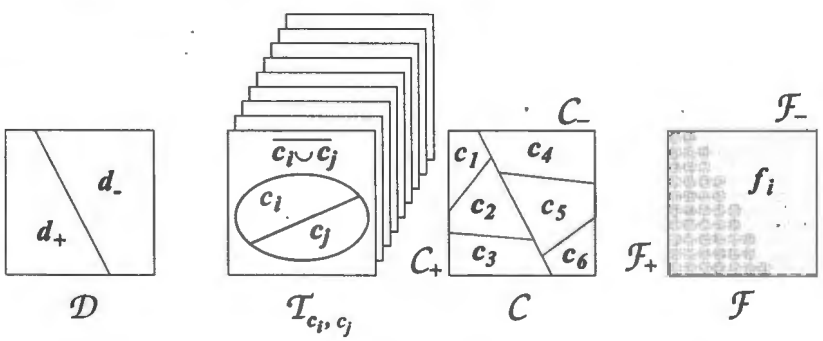
Selekcja wstępna ma charakter klasyfikacji binarnej. Polega na odróżnieniu danych *akceptowanych*, przeznaczonych do dalszego złożonego przetwarzania w ramach realizacji eksperymentu, od danych *dyskwalifikowanych*. Jednak ze względu na wysoką wymiarowość przestrzeni reprezentacji i znaczne zróżnicowanie wewnętrzne obydwu klas wskazane jest ich rozwarstwienie poprzez wyodrębnienie w nich pewnych podklas.

Na rozwarstwioną przestrzeń klas złoży się $l > 2$ rozłącznych *klas diagnostycznych* $C = \{c_1, c_2, \ldots, c_l\}$. Zbiór klas C jest tak zdefiniowany, że każde rozstrzygnięcie dokonane w jego ramach, implikuje jednoznaczną decyzję w zbiorze $D = \{akceptacja, dyskwalifikacja\}$. Zbiór klas diagnostycznych C stanowi więc uszczegółowienie zbioru D. Stworzenie klasyfikacji diagnostycznej wymaga posia-

dania pewnej obiektywnej, lub postulowanej wiedzy a priori o selekcjonowanych danych i strukturze ich populacji. W rzeczywistych zadaniach z rozważanego tu obszaru zwykle dostępne są modele teoretyczne, wprowadzające szczegółową kategoryzację badanych obiektów; oznaczmy taki zbiór kategorii jako $F = \{f_1, f_2, \ldots, f_n\}$. Dane pochodzące z eksperymentów mogą jednak nie dostarczać przesłanek wystarczających, by móc na ich podstawie dokonać rozróżnień tak subtelnych jak przewiduje to kategoryzacja F. Klasyfikację C należy wówczas zdefiniować jako uogólnienie F. W sformułowaniu właściwym dla teorii Dempstera-Shafera, zbiory D i F wyznaczają zgodne przestrzenie rozważań (ang. *compatible frames of discernment*) o skrajnych skalach szczegółowości, co objaśnia rys. 1.

Zestaw klas diagnostycznych jest parametrem strukturalnym procedury selekcji, i może podlegać optymalizacji. Należy go dobrać tak, by zredukować koszt rozwiązania jednostkowych zadań obliczeniowych, starając się przy tym wpłynąć korzystnie na efektywność mających powstać klasyfikatorów[1].

Pierwszy cel można osiągnąć wprowadzając dostatecznie dużą liczbę klas diagnostycznych; drugi, odpowiednio je dobierając.



Rysunek 1: Rodzina zgodnych przestrzeni rozważań rozpatrywanych w ramach proponowanej procedury selekcji. Przestrzeń D wynika bezpośrednio z celu selekcji, czyli dokonania akceptacji (klasa d_+) albo dyskwalifikacji (klasa d_-) danych. Elementy wyróżnione w przestrzeniach $T_{ci,cj}$ reprezentują odpowiedzi klasyfikatorów binarnych (testów diagnostycznych), niezależnie analizujących dane. Złożenie wyników testów umożliwia wnioskowanie odnoszące się do wieloklasowego podziału populacji w przestrzeni C. Przestrzeń C stanowi minimalne wspólne uszczegółowienie prezentowanej rodziny przestrzeni rozważań. Klasa przypisana w ramach C ma swój jednoznaczny odpowiednik w D. Podstawą wszystkich podziałów jest przestrzeń F, rozróżniająca pojęcia odnoszące się do teoretycznego opisu selekcjonowanych danych

[1] Motywacją posłużenia się zespołem klasyfikatorów jest zmniejszenie złożoności jednostkowych zadań obliczeniowych. Podejście takie można traktować jednak ogólniej, jako metodę pokonania ograniczeń właściwych dla pojedynczych klasyfikatorów (Marciniak, 2002).

Naszym celem jest zaproponowanie takiej metody konstrukcji klasyfikatora operującego na zbiorze klas C, która ograniczy liczbę elementów najliczniejszego zbioru danych wymaganego w procesie uczenia. Osiągamy to stosując metodę dekompozycji zadania klasyfikacji, określaną mianem „1-vs-1" (Hsu i Lin, 2002). Polega ona na przygotowaniu dla każdej z par klas c_i, $c_j \in C$, $i \neq j$ osobnego klasyfikatora (testu) binarnego T_{c_i,c_j}, uczonego wyłącznie na przykładach z tej pary klas. Na pełen zestaw klasyfikatorów binarnych w podejściu „1-vs-1" składa się więc $l(l-1)/2$ testów, gdzie l to liczność rozważanego zbioru klas C. Zakładając, że udział reprezentantów poszczególnych klas w zbiorze uczącym, złożonym z m przykładów, wynosiłby średnio m/l, optymalizacja pojedynczego testu w układzie „1-vs-1", wymagałaby uwzględnienia $2m/l$ przykładów. Zatem, gdy złożoność obliczeniowa procedury uczenia testu zależy kwadratowo od liczby przykładów, koszt optymalizacji pojedynczego testu stanowi $4/l^2$ kosztu optymalizacji klasyfikatora jednolitego. Łączny nakład obliczeń związanych z przygotowaniem wszystkich $l(l-1)/2$ testów „1-vs-1" jest rzędu $O(m^2)$, i nie zależy od liczby klas l. Jest on więc porównywalny z kosztem zadania skonstruowania klasyfikatora jednolitego, jednak koszt pojedynczych zadań optymalizacyjnych zostaje ograniczony.

Proponowana procedura selekcji wstępnej sprowadza się więc w dużym stopniu do zadania klasyfikacji z wieloma klasami, realizowanego przy użyciu zespołu klasyfikatorów binarnych. Odpowiednia procedura scalająca odpowiedzi klasyfikatorów binarnych może powstać jako adaptacja metod opracowanych dla standardowych zagadnień klasyfikacji, z tą różnicą, że w rozważanym tu zadaniu nie jest koniecznym wskazanie konkretnej klasy $c^* \in C$. Wystarczającym jest zaliczenie obiektu do pewnego podzbioru klas diagnostycznych C^*, który zawiera się w podzbiorze C_+, złożonym z wszystkich klas poszukiwanych (odpowiadającym klasie „*akceptacja*" ze zbioru D), albo jego dopełnieniu $C_- = C \setminus C_+$. Klasyfikatory binarne składające się na zespół będą nazywane *testami diagnostycznymi*.

Przyjmuje się, iż selekcja wstępna jest realizowana w ten sposób, że dane zbierane w trakcie eksperymentu poddawane są sprawdzeniu z użyciem każdego z testów diagnostycznych, a uzyskane wyniki są agregowane w celu podjęcia ostatecznej decyzji. Tę agregację (fuzję) wyników testów diagnostycznych przeprowadza się z użyciem teorii Dempstera-Shafera, rozumując jak opisano poniżej.

Wynik testu T_{c_i,c_j} jedynie uprawdopodabnia prawdziwość tezy, że badany obiekt należy do jednej spośród klas c_i i c_j. W rzeczywistości obiekt może być reprezentantem drugiej z nich, albo jeszcze innej klasy, różnej zarówno od c_i, jaki

i od cj . Wobec tej niejednoznaczności, przyjmujemy że każdy z testów diagnostycznych pozwala określić, dla podanych na wejściu danych s, w jakim stopniu wykluczona jest ich „przynależność" do każdej z dwóch klas ci i cj . Natomiast ignorancję co do faktycznej przynależności danej do którejś z klas C wyrażamy trywialnym stwierdzeniem, iż $s \in C$. Traktujemy więc wynik każdego testu $T_{ci,cj}$ jako źródło dwóch *przesłanek* (ang. *bodies of evidence*), które muszą by rozpatrywane razem. Na każdą z nich składają się dwa stwierdzenia[2].

Przesłanka 1 („¬j") obejmuje dwa następujące stwierdzenia:

1. s nie należy do klasy c_j ($s \in \overline{\{c_j\}}$), oraz
2. $s \in C$

Przesłanka 2 („¬i") obejmuje dwa następujące stwierdzenia:

1. s nie należy do klasy c_i ($s \in \overline{\{c_i\}}$), oraz
2. $s \in C$

Przekonanie na temat przynależności danych s do poszczególnych klas w zbiorze C, w ogólności do jednej z klas tworzących dowolny podzbiór $A \subseteq C$, wynikające niezależnie z każdej z tych dwóch przesłanek, wyrazimy formalnie poprzez dwie funkcje przekonania Bel, będące prostymi funkcjami wsparcia:

1. $\mathrm{Bel}_{\neg j}(s \in A)$
2. $\mathrm{Bel}_{\neg i}(s \in A)$

Odpowiadające tym funkcjom *bazowe rozkłady prawdopodobieństwa* (ang. *basic probability assignment; basic belief assignment; mass assignment*) oznaczymy jako: $m_{\neg j}$ oraz $m_{\neg i}$. Mają one następującą postać:

$$m_{\neg k}(s \in \overline{\{c_k\}}) = \mathrm{Bel}_{\neg k}(s \in \overline{\{c_k\}})$$

$$m_{\neg k}(s \in C) = 1 - \mathrm{Bel}_{\neg k}(s \in \overline{\{c_k\}}) \qquad (1)$$

$$m_{\neg k}(s \in A) = 0 \text{ dla wszystkich pozostałych } A \subset C,$$

gdzie $k = i$ lub odpowiednio $k = j$.

[2] Zapis s A oznacza tu i w dalszej części artykułu, że s należy do jednej z klas podzbioru A _ C.

Obie przesłanki wynikające z wyniku testu $T_{ci,cj}$, rozpatrywane łącznie, implikują bazowy rozkład prawdopodobieństwa dany sumą ortogonalną $m_{\neg j}$ i

$$m_{\neg i} : mT_{ci,c} = m_{\neg j} \oplus m_{\neg i}, \tag{2}$$

$mT_{ci,cj}$ niezerowe wartości może przybrać wyłącznie dla stwierdzeń ze zbioru

$$T_{ci,cj} : T_{ci,cj} = \{s \in \overline{\{c_i\}}, s \in \overline{\{c_j\}}, s \in \overline{\{c_i \cup c_j\}}, s \in C\}. \tag{3}$$

Interpretując wyniki testu $T_{ci,cj}$ nie stwierdza się, iż jeśli obiekt nie należy do klasy c_i, to musi należeć do klasy c_j, tym samym nie żąda się spełnienia warunku $Bel_{\neg i}(s \in \overline{\{c_i\}}) + Bel_{\neg j}(s \in \overline{\{c_j\}}) = 1$.

W proponowanym podejściu, stopnie przekonań $Bel_{\neg k}(s \in \overline{\{c_k\}})$ przyjmowane są jako równe wartości pewnej dobieranej empirycznie funkcji $bel_{\neg k}(s)$ określonej na zbiorze cech obiektu s.

Podanie definicji $bel_{\neg k}(s)$ sprowadza się do przyjęcia odpowiedniej funkcji $f(s)$, powiązanej z wielkościami wyznaczanymi podczas wykonywania algorytmu testu $T_{ci,cj}$, i decydującymi o jego wyniku. W zależności od konkretnego rozwiązania, rolę tę może pełni funkcja dyskryminacyjna testu, albo wyznaczana w nim ranga. Wartości $f(s)$ stają się dziedziną kolejnej funkcji $T(f(s))$, posiadającej własności formalne stopnia przekonania:

$$bel_{\neg k}(s) = T(f(s)). \tag{4}$$

Funkcje $bel_{\neg i}(s)$ oraz $bel_{\neg j}(s)$, muszą być konstruowane niezależnie od siebie. W proponowanym podejściu za podstawę do ich konstrukcji służą odpowiednio podzbiory przykładów obserwacji $\{s : s \in \{c_i\}\}$ i $\{s : s \in \{c_j\}\}$.

Istnieje pewna dowolność wyboru funkcji $f(s)$ i $T(f(s))$. Na wzór regresyjnych modeli klasyfikacji, można tu zastosować zależność logistyczną (Schölkopf, 2002):

$$T(f(s)) = 1/[1 + \exp(f(s))]. \tag{5}$$

3. Fuzja wyników zespołu testów diagnostycznych

Wyniki testów diagnostycznych $T_{ci,cj}$ tworzących zespół $\{T_1, T_2, \ldots, T_k\}$, które reprezentowane są przez bazowy rozkład prawdopodobieństwa (2), mogą zostać połączone według reguły Dempstera, drogą obliczenia sumy ortogonalnej $m_{\oplus} = m_{T_1} \oplus m_{T_2} \oplus \cdots \oplus m_{T_k}$. Wzajemna *zależność znaczeniowa* przestrzeni

rozważań $T_{ci,cj}$ 3, w których testy dokonują rozróżnień, sprawia, że rozkłady m_{Ti} są kombinowalne.

Rozstrzygnięcia końcowe *D*, o akceptacji, albo dyskwalifikacji danej s mogą zapadać na podstawie dyskryminacji wartości pewnej funkcji wartości bazowego rozkładu prawdopodobieństwa $m_{\oplus}$, przypisanych wybranym rodzinom podzbiorów klas diagnostycznych. Prostym przykładem takiej funkcji może być stopień domniemania (ang. *plausibility measure*), Pl, obliczany dla zbioru C+ (zdefiniowanego na s. 109). Tak więc ustala się pewną wartość progową Pl_{min} i dana s jest klasyfikowana jako „*akceptowana*", jeśli:

$$Pl(s \in \{C+\}) = \sum_{B \subseteq C: B \cap C_+ \neq 0} m_{\oplus}(s \in B) > Pl_{min}$$

Warto zauważyć, że spośród ogółu testów możliwych do skonstruowania w ramach klasyfikacji diagnostycznej C, nie wszystkie dostarczą informacji równie znaczącej dla zadania wyselekcjonowania z danych obserwacyjnych reprezentantów podzbioru C+. Różnice w znaczeniu poszczególnych testów w zespole dla podjęcia decyzji końcowej można objaśnić na przykładzie przedstawionym na rys. 1, w którym wynik wykluczający przynależność obiektu do jednej z klas podzbioru $C+ = \{c_1, c_2, c_3\}$ przemawia za dyskwalifikacją (*d*–), a wynik eliminujący obiekt z podzbioru $C- = \{c_4, c_5, c_6\}$ za akceptacją (*d*+).

Test $T_{c1,c2}$, rozróżniając klasy wewnątrz podzbioru C+, nie pomaga wprost w podjęciu decyzji końcowej: *d*+ albo *d*–, podobnie wynik testu $T_{c4,c5}$, operującego wewnątrz podzbioru C–. Inaczej natomiast test $T_{c1,c4}$, preferując klasę $c_1 \in C+$ albo $c_4 \in C-$, wprost pomaga wybrać jedną z decyzji końcowych.

Bazowy rozkład prawdopodobieństwa $m_{\oplus} = m_{T1} \oplus m_{T2} \oplus \cdots \oplus m_{Tk}$ przypisuje zbiorom $B \subseteq C$ wartości proporcjonalne do sumy iloczynów wartości składowych bazowych rozkładów prawdopodobieństwa m_{Ti}:

$$m_{\oplus}(B) \propto$$

$$\sum_{Bi_1, Bi_2, \ldots, Bi_k : Bi_1 \cap Bi_2 \cap \ldots \cap Bi_k = B} m_{T1}(B_{i1}) m_{T2}(B_{i2}) \ldots m_{Tk}(B_{ik}).$$

By wyznaczyć rozkład jej wartości koniecznym jest zbadanie przecięć *k* zbiorów $B_{i1} \cap B_{i2} \cap \ldots \cap B_{ik}$, przy wyczerpaniu wszystkich możliwości wyboru każdego z B_{ij} spośród liczącego 2^l elementów zbioru 2^C. W ogólności złożoność takiej operacji jest rzędu $O(2^{lk})$. Jednak w przypadku kombinowania funkcji reprezentujących wyniki testów $T_{ci,cj}$, jedynie dla zbiorów

$$B_{ij} \in \{\overline{\{c_i\}},\ \overline{\{c_j\}},\ \overline{\{c_i \cup c_j\}}, C\} \tag{6}$$

może zachodzić $m_{\tau ci,cj}(B_{ij}) \neq 0$. Zatem wiadomym jest, które składniki powyższej sumy mogą być niezerowe, i przy wyznaczaniu wartości rozkładu $m_{\oplus}$ wystarczy zbadać przecięcia zbiorów B_{ij}, wybieranych z 4-elementowych rodzin (por. (6)), co efektywnie redukuje złożoność obliczeń do rzędu $0(4^k)$.

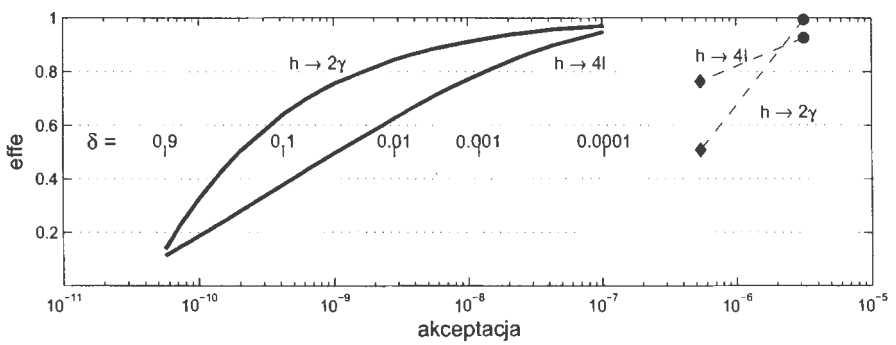
4. Eksperymenty obliczeniowe

Zaproponowany algorytm przetestowano w symulacji numerycznej eksperymentu z dziedziny fizyki cząstek elementarnych, wzorowanego na CMS CMS (2003), w którym poszukiwano danych pomiarowych świadczących o wystąpieniu 2-fotonowego lub 4-leptonowego rozpadu cząstki Higgsa, wytworzonej podczas oddziaływania wysokoenergetycznych protonów. W rzeczywistym eksperymencie względna częstość występowania poszukiwanych danych wynosi $< 1/10^8$, i wymaga się by odsetek danych zaakceptowanych do dalszej analizy nie przekraczał $1/10^6$. W symulacji numerycznej dane reprezentowane były jako wektory w przestrzeni 1152 wymiarowej. Do selekcji każdego z dwóch poszukiwanych rodzajów danych (rozpad 2-fotonowy lub 4-leptonowy) przygotowano osobną procedurę. W każdej z nich, na podstawie modelu teoretycznego zjawisk fizycznych wydzielono 7 klas diagnostycznych, $C = \{c_1, \ldots, c_6, c_+\}$, przy czym zbiór C_+ tworzyła klasa c_+. Użyto zespołów składających się z 6 testów diagnostycznych w układzie „c_+ vs c_k", $k = 1, \ldots, 6$.

Testy diagnostyczne $T_{ci,cj}$ skonstruowano jako nieliniowe klasyfikatory SVM (Schölkopf, 2002), uczone na podstawie przykładów obserwacji z klas c_i i c_j. Do określenia stopni przekonania $\mathrm{Bel}_{\neg ck}(s \in \overline{\{c_k\}}) = bel_{\neg k}(s)$ wykorzystano wartość $\rho(s)$, wyznaczaną standardowo przy wykonywaniu algorytmu klasyfikatora SVM. Wartość $\rho(s)$ równa jest odległości obrazu wektora danych $\phi(s)$, w docelowej przestrzeni (ang. *feature space*) stosowanej przez SVM, od głównej hiperpłaszczyzny separującej maszyn SVM, w tej przestrzeni. Funkcję $bel_{\neg k}(s)$ definiuje się zgodnie z (4) i (5), przy czym

$$f(s) = \beta(\rho(s) - \alpha)$$

gdzie α i β są parametrami dobieranymi tak, żeby dopasować $bel_{\neg k}(s)$ do unormowanych histogramów dystrybuanty empirycznej warunkowych rozkładów zmiennej $\rho(s \mid c_k)$, uzyskanych na podstawie zbiorów uczących.



Rysunek 2: Efektywność (effe) (ang. *True Positive Rate*) wykrywania w danych wejściowych przypadków wystąpień 2-fotonowych (h → 2) i 4-leptonowych (h → 4l) rozpadów cząstek Higgsa, w zależności od poziomu akceptacji całkowitego strumienia danych wejściowych (akceptacja). Krzywe ciągłe, sparametryzowane wartością progową funkcji dyskryminacyjnej (_), odpowiadają selekcji dokonywanej na podstawie sumy logicznej odpowiedzi klasyfikatora wytrenowanego do wykrywania rozpadów 2-fotonowych oraz klasyfikatora wykrywającego rozpady 4-leptonowe, pracujących z tą samą wartością progową. Punkty połączone liniami przerywanymi opisują wyniki selekcji prowadzonej przy użyciu procedury stworzonej na podstawie modelu teoretycznego oddziaływania protonów (Gościło, 2000), odpowiednio jej wersje: o wyższej (koła) i niższej efektywności (romby)

Ostateczna klasyfikacja danych jako „akceptowanych" bądź „dyskwalifikowanych" dokonywana jest na podstawie dyskryminacji wartości funkcji zdefiniowanej jako stosunek wartości sum mas prawdopodobieństwa m⊕ przypisanych wybranym rodzinom podzbiorów klas diagnostycznych. Przetestowano różne warianty składu tych rodzin podzbiorów.

Parametry użytkowe najlepszego z uzyskanych wariantów procedury selekcji wstępnej, oszacowane na pochodzących z symulacji danych testowych, przedstawia rys. 2. Można zauważyć, że z użyciem proponowanej w artykule metody uzyskano lepsze wyniki, niż w przypadku metody selekcji wywiedzionej bezpośrednio z modelu teoretycznego oddziaływań protonów.

5. Uwagi końcowe

W artykule zaproponowano efektywną metodę konstruowania klasyfikatorów przeznaczonych do selekcji wstępnej danych gromadzonych w trakcie eksperymentów fizycznych. Należy zaznaczyć, że eksperymenty, z myślą o których opracowano opisaną metodę, mają charakter unikalny i dostępność materiału porównawczego jest ograniczona. W przeprowadzonym eksperymencie obliczeniowym, wyniki działania klasyfikatora skonstruowanego proponowana ˛ metodą, są

lepsze od wyników uzyskanych z pomocą klasyfikatorów o algorytmach podanych w literaturze (Gościło, 2000).

W ramach ogólnego podejścia zaproponowano oryginalne rozwiązanie fuzji zespołu klasyfikatorów przy użyciu metod teorii Dempstera-Shafera. W literaturze znane są podejścia tego typu (por. (Rogova, 1994)), ale w proponowanym rozwiązaniu przyjęto oryginalną interpretację wyników klasyfikatorów w ramach dekompozycji zadania klasyfikacji zgodnie z podejściem „1-vs-1".

Proponowane rozwiązanie jest wysoce sparametryzowane. Dalsze prace koncentrować się będą wokół bardziej szczegółowego przebadania efektywności uzyskanego klasyfikatora w zależności od przyjętych wartości parametrów

Literatura

[1]. CMS Collaboration (2003): Compact Muon Solenoid Experiment, CERN, http://cms.web.cern.ch/cms/Media/Publications/Brochures/.

[2]. Gościło Ł. (2000): Symulacja odpowiedzi wielostopniowego systemu wyzwalania detektora CMS, praca magisterska wykonana w Instytucie Fizyki Doświadczalnej Wydziału Fizyki Uniwersytetu Warszawskiego, Warszawa.

[3]. Hsu C.W, Lin C.J. (2002): A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2), 415-425.

[4]. Marciniak A., Korbicz J. (2002): Metody rozpoznawania obrazów w diagnostyce. W: Diagnostyka procesów, Korbicz J., Kościelny J.M., Kowalczuk Z., Cholewa W. (red.) WNT, Warszawa.

[5]. Rogova G. (1994): Combining the results of several neural-network classifiers. *Neural Networks*, 7(5), 777-781.

[6]. Schölkopf B., Smola A.J. (2002): *Learning with Kernels*. MIT Press, Cambridge.

[7]. Shafer G. (1976): *A mathematical theory of evidence*. Princeton University Press, Princeton.

[8]. Sobczak W., Malina W. (1985): *Metody selekcji i redukcji informacji*, wyd. 2. WNT, Warszawa.

ISBN 9788389475220