Autoreferat
przedstawiający opis dorobku
i osiągnięć naukowych
habilitanta dra inż. Przemysława Klęska

1 Imię i nazwisko

Przemysław Klęsk (urodzony 9 sierpnia 1977 r. w Koszalinie)

2 Posiadane dyplomy, stopnie naukowe

Oświadczam, że posiadam stopień doktora nauk technicznych nadany mi przez Radę Wydziału Informatyki Politechniki Szczecińskiej dnia 17 maja 2005 r. w Szczecinie. Tytuł pracy doktorskiej: *„Metoda nadawania pożądanych własności ekstrapolacyjnych neuronowym i rozmytym modelom systemów wielowymiarowych"*. Promotor pracy: prof. dr hab inż. Andrzej Piegat. Praca obroniona z wyróżnieniem.

3 Informacje o dotychczasowym zatrudnieniu w jednostkach naukowych

Od 1 października 2005 r. pozostaję zatrudniony w Katedrze Metod Sztucznej Inteligencji i Matematyki Stosowanej[1] przy Wydziale Informatyki Zachodniopomorskiego Uniwersytetu Technologicznego w Szczecinie, ul. Żołnierska 49, 71-210 Szczecin (wcześniej: Politechnika Szczecińska).

4 Wskazanie osiągnięcia

Jako osiągnięcie (wynikające z art. 16 ust. 2 ustawy z dnia 14 marca 2003 r. o stopniach naukowych i tytule naukowym oraz o stopniach i tytule w zakresie sztuki) wskazuję **jednotematyczny cykl publikacji pt. „Zdolność do uogólniania w uczeniu maszynowym"**. Na cykl składa się sześć publikacji (wymienionych poniżej), które widnieją w pełnym spisie publikacji pod numerami: 1, 2, 3, 4, 8 w załączniku „Wykaz opublikowanych prac naukowych" oraz dodatkowo publikacja nr 1 z punktu 4 tego załącznika. Główne rezultaty habilitanta pochodzące z tych prac są omówione w niniejszym autoreferacie zgodnie z poniższym porządkiem.

4a Jednotematyczny cykl publikacji

Klęsk, P. (2011), A Relationship Between Cross-Validation and Vapnik Bounds on Generalization of Learning Machines, *in* 'Proceedings of the 3-rd International Conference on Agents and Artificial Intelligence — ICAART 2011', Vol. 1, SciTePress, Rome, Italy, pp. 5–17.

Klęsk, P. (2010*b*), 'Probabilities of discrepancy between minima of cross-validation, Vapnik bounds and true risks', *International Journal of Applied Mathematics and Computer Science* **20**(3), 525–544. Zielona Góra, Poland.

Klęsk, P. (2010*a*), 'A comparison of certain generalization bounds of learning machines for practical applications', *Metody Informatyki Stosowanej* **2**(24), 35–45. Polska Akadamia Nauk oddział w Gdańsku, Szczecin, Poland.

[1] Wcześniejsza nazwa: Instytut Sztucznej Inteligencji i Robotyki.

Klęsk, P. i Korzeń, M. (2011), 'Sets of approximating functions with finite Vapnik-Chervonenkis dimension for nearest neighbors algorithms', *Pattern Recognition Letters* **32**(14), 1182–1893. Elsevier, New York, USA.

Klęsk, P. (2012), Probabilistic Estimation of Vapnik-Chervonenkis Dimension, *in* 'Proceedings of the 4-th International Conference on Agents and Artificial Intelligence — ICAART 2012', SciTePress, Vilamoura, Portugal.

Korzeń, M. i **Klęsk, P.** (2008), Maximal Margin Estimation with Perceptron-like Algorithm, Vol. 5097/2008 of *Lecture Notes in Artificial Intelligence*, Springer, Berlin / Heidelberg, Germany, pp. 597–608. 9th International Conference on Artificial Intelligence and Soft-Computing — ICAISC 2008, Zakopane, Poland

Oświadczenia o wkładzie współautorów

W przypadku dwóch publikacji (spośród wskazanych sześciu publikacji stanowiących jednotematyczny cykl) współautorem jest dr inż. Marcin Korzeń (zatrudniony w tej samej jednostce naukowej). Poniżej przedstawiono oświadczenia o indywidualnym wkładzie współautorów w te publikacje.

1. **Publikacja (Klęsk i Korzeń, 2011)** p.t. "Sets of approximating functions with finite Vapnik-Chervonenkis dimension for nearest neighbors algorithms".

   Elementy udziału P. Klęska (60% całości): (a) pomysł na sformułowanie algorytmu k-najbliższych sąsiadów jako algorytmu ze stałym procentem α-najbliższych sąsiadów, (b) sposób zdefiniowania zbioru funkcji aproksymujących dla algorytmu α-NN* i ich stopni swobody, (c) przypuszczenie o wymiarze VC równym $\lfloor 2/\alpha \rfloor$, (d) dowód lematu (Lemma 2) o gwarantowanej liczbie dychotomii, (e) wykorzystanie wymiaru VC algorytmu α-NN* w procedurze wyboru złożoności modelu (SRM), (f) eksperymenty z rozpoznawaniem wzorca szachownica, (g) eksperymenty porównawcze z siecią neuronową RBF.

   Elementy udziału M. Korzenia (40% całości): (a) dowód lematu (Lemma 3) o niemożliwości wszystkich dychotomii, (b) uwagi o klasyfikatorach 'instance-based', (c) dodatkowy algorytm uczenia dla α-NN* oparty o klasteryzację.

   Elementy opracowane wspólnie: (a) sformułowanie głównego twierdzenia (Theorem 1) o skończonym wymiarze VC dla zbioru funkcji związanego z algorytmem α-NN*. (b) udowodnienie konsekwencji 1 (Corollary 1) o gwarantowanej liczbie dychotomii.

   Podpisy oświadczających: ..... (dr inż. P. Klęsk) ..... (dr inż. M. Korzeń)

2. **Publikacja (Korzeń i Klęsk, 2008)** p.t. "Maximal Margin Estimation with Perceptron-like Algorithm".

   Elementy udziału M. Korzenia (70% całości): (a) pomysł na modyfikację algorytmu perceptronu poprzez wprowadzenie marginesu separacji, (b) definicja γ-separowalności i dowód istnienia marginesu separacji (w szczególności ujemnego) dla każdego zbioru danych, (c) dowód zbieżności zmodyfikowanego algorytmu przy ustaleniu dodatniego marginesu, (d) strategie zatrzymywania algorytmu gdy optymalny nieznany margines jest mniejszy niż margines zadany, (e) algorytm

varying-margin, (f) eksperymenty porównawcze z innymi klasyfikatorami (w tym wykrywanie marginesu separacji w danych).

Elementy udziału P. Klęska (30% całości): (a) algorytm *soft-decrease margin*, (b) eksperymenty porównawcze algorytmów *varying-margin* i *soft-decrease margin*.

Elementy opracowane wspólnie: (a) praca nad dowodem dla przypadku ujemnego marginesu, (b) związki proponowanego algorytmu z algorytmem SVM.

Podpisy oświadczających: ....

(dr inż. P. Klęsk) (dr inż. M. Korzeń)

4b Omówienie celu naukowego i osiągniętych wyników

Wprowadzenie

Głównym obszarem zainteresowań naukowych habilitanta są **algorytmy uczenia maszynowego**. Typowym postawieniem problemu związanym z tymi algorytmami jest uczenie się z danych, innymi słowy — budowanie modelu na podstawie danych, w tzw. *sytuacji obserwacyjnej* (ang. *observational setting*). Jest to sytuacja częsta w rzeczywistości. Jesteśmy biernymi obserwatorami pewnego zjawiska, odnotowujemy pochodzące z niego dane, natomiast nie znamy mechanizmu rządzącego tymże zjawiskiem. Mówiąc ściślej nie znamy łącznego rozkładu prawdopodobieństwa, według którego objawiają się obserwowane wielkości.

Jak można zauważyć, uczenie się z danych stało się w ostatniej dekadzie na świecie bardzo popularne. Widoczne jest to zwłaszcza w takich dziedzinach jak np.: biotechnologia, medycyna, ekonomia, socjologia, gdzie pojawiło się dużo praktycznych aplikacji opartych na gromadzonych zbiorach danych. Istnieją pewne algorytmy, które sprawdzają się szczególnie dobrze w praktyce i są obecnie uznawane za tzw. *state-of-art* m.in.: maszyny SVM (Vapnik, 1998), drzewa decyzyjne C4.5 (Quinlam, 1993), krzywe sklejane MARS (Friedman, 1991), perceptrony wielowarstwowe (Cybenko, 1989; Rumelhart, Hinton i Williams, 1986).

Habilitant koncentruje swoją uwagę badawczą tylko po części na konkretnych algorytmach, natomiast w większej mierze na pewnych matematycznych własnościach / problemach związanych z uczeniem maszynowym w sensie szerszym, takich jak:

- **zdolność maszyn uczących się do uogólniania,**
- **ograniczenia liczbowe na błąd prawdziwy,**
- **techniki wyboru złożoności modelu,**
- **ocena bogatości zbiorów funkcji stosowanych do uczenia,**
- **związki pomiędzy testowaniem o uogólnianiem.**

Od strony formalnej, naturalnym narzędziem do prowadzenia badań w tym kierunku jest *Statystyczna Teoria Uczenia* (ang. *Statistical Learning Theory*) zapoczątkowana przez Vladimira Vapnika (Vapnik, 1998; Vapnik, 1995) oraz model matematyczny PAC (ang. *probably approximately correct*) (Valiant, 1984). W tym ujęciu, na daną maszynę uczącą się patrzymy jak na parę: (1) zbiór funkcji matematycznych, który ma ona do dyspozycji oraz (2) algorytm uczący, który mówi, w jaki sposób należy wybrać jedną

funkcję z tego zbioru. Zadanie uczenia się z danych dobrze jest wówczas postawić jako problem, który należy rozwiązać z zadaną z góry (ϵ, δ)-precyzją. Oznacza to, że algorytm uczący będzie wybierał funkcję, która popełnia błąd prawdziwy nie gorszy niż o ϵ od błędu najlepszej funkcji możliwej do osiągnięcia w przyjętym zbiorze, i fakt ten będzie miał miejsce z prawdopodobieństwem przynajmniej $1 - \delta$.

Jak już określono wcześniej, w dorobku habilitanta można wskazać **jednotematyczny cykl publikacji** (sześć publikacji) dotyczący zdolności do uogólniania w uczeniu maszynowym. Jako „tło" w punkcie 4b.1 przedstawiono niektóre ważne rezultaty światowe z tej dziedziny. Główne rezultaty z prac habilitanta zostały przedstawione w punkcie 4b.2. Pełne wersje tych prac zostały załączone do wniosku habilitacyjnego.

4b.1 Notacja, podstawowe pojęcia, niektóre znane rezultaty

W niniejszym autoreferacie przyjęto notację, którą stosują m.in. Anthony i Bartlett (2009). W artykułach habilitanta, do których będą następowały odniesienia, bywa nieco częściej używana notacja zbliżona do tej, którą stosują Vapnik (1998) czy też Cherkassky i Mulier (1998). Dla ułatwienia ewentualnej lektury treści samych artykułów, istotne różnice notacyjne będą sygnalizowane.

Niech

$$\mathbf{z} = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_m, y_m)\}. \tag{1}$$

oznacza **próbę** o rozmiarze m, tj. zbiór par $(\mathbf{x}, y)$ czerpanych z pewnego *nieznanego*, ale *stałego* rozkładu prawdopodobieństwa P. Rozkład P reprezentuje dane zjawisko, które jest przedmiotem uczenia. Punkty danych czerpane są w sposób i.i.d. (ang. *independent, identically distributed*) i tym samym możemy myśleć o produktowym rozkładzie P^m, z którego pochodzi cała próba. Jeżeli chodzi o dziedziny, niech w ogólności $\mathbf{x} \in \mathbf{X} \subset \mathbb{R}^d$ oraz $y \in Y$, gdzie w zależności od rodzaju zadania dziedzinę Y będzie stanowił pewien zbiór skończony (zadanie klasyfikacji), lub zbiór $\mathbb{R}$ (zadanie estymacji funkcji regresji).

Niech

$$F = \left\{f\right\}, \tag{2}$$

gdzie $f\colon \mathbf{X} \to Y$, oznacza **zbiór funkcji**, który maszyna ucząca się ma do dyspozycji. Za pomocą pewnej wybranej funkcji z tego zbioru będziemy chcieli przybliżać badane zjawisko. Zbiór F bywa w literaturze nazywany również zbiorem hipotez[2].

Pojęcie **zdolności do uogólniania** dla pewnej ustalonej funkcji f można utożsamiać z liczbową wartością **błędu prawdziwego**[3] (ang. *true error*) popełnianego przez tę funkcję. Chodzi tu o błąd policzony w sposób dokładny jako wartość oczekiwaną względem rozkładu P. Dla zadania klasyfikacji błąd prawdziwy definiujemy jako:

$$\mathrm{er}_P(f) = \int_{\mathbf{x}\in\mathbf{X}} \sum_{y\in Y} [f(\mathbf{x}) \neq y] \underbrace{p(\mathbf{x})P(y|\mathbf{x})\mathbf{dx}}_{dP(\mathbf{x},y)}, \tag{3}$$

gdzie notacja $[\cdot]$ jest funkcją wskaźnikową, przyjmującą 1 gdy zdanie będące argumentem jest prawdziwe i 0 w przeciwnym razie. Jak można zauważyć, liczbowy sens er_P to *prawdopodobieństwo błędnego sklasyfikowania* losowej pary $(\mathbf{x}, y)$ zaczerpniętej z P. Dla zadania estymacji funkcji regresji błąd prawdziwy

[2] W niektórych artykułach habilitanta dla zbioru funkcji maszyny uczącej się przyjęta jest notacja $\{f(\mathbf{x}, \omega)\}_{\omega\in\Omega}$, gdzie Ω oznacza przestrzeń wartości parametrów tego zbioru. Tym samym małe ω można rozumieć, jak indeks konkretnej funkcji w zbiorze. Innymi słowy jest to notacja z wyraźnym „wyłuszczeniem" parametryzacji zawartej w przyjętej postaci funkcyjnej.

[3] Wielkość ta bywa też nazywana *ryzykiem prawdziwym* (ang. *true risk*) np. u Vapnika.

definiujemy zwykle[4] jako:

$$\mathrm{er}_P(f) = \int_{\mathbf{x}\in\mathbf{X}} \int_{y\in Y} (f(\mathbf{x}) - y)^2 \underbrace{p(\mathbf{x})p(y|\mathbf{x})dy\mathbf{dx}}_{dP(\mathbf{x},y)}. \tag{4}$$

Liczbowy sens er_P to w tym przypadku *oczekiwany kwadrat odchyłki* pomiędzy $f(\mathbf{x})$ a y.

Należy zaznaczyć, że błąd prawdziwy oczywiście *nie* jest w praktyce możliwy do obliczenia, ponieważ nieznany jest rozkład P, a do naszej dyspozycji jest tylko zaobserwowana próba pochodząca z tego rozkładu. Interesującym natomiast jest to, że w ramach teorii SLT istnieją różne techniki pozwalające na szacowanie tej nieznanej wartości.

Podstawową wielkością, która jest wyliczana i pojawia się w każdym praktycznym eksperymencie, jest **błąd na próbie**[5] (ang. *sample error*). Dla zadania klasyfikacji błąd na próbie $\widehat{\mathrm{er}}_{\mathbf{z}}$ obliczamy jako

$$\widehat{\mathrm{er}}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}[f(\mathbf{x}_i) \neq y_i], \tag{5}$$

natomiast dla zadania estymacji funkcji regresji jako

$$\widehat{\mathrm{er}}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}(f(\mathbf{x}_i) - y_i)^2. \tag{6}$$

Liczbowy sens (5) i (6) to odpowiednio *częstość błędnej klasyfikacji* oraz *średni kwadrat odchyłki*.

Algorytm uczący L wybiera ze zbioru F jedną funkcję $\widehat{f}$ mając na uwadze zaobserwowane dane, starając się, aby wybrana funkcja minimalizowała błąd na próbie, tj. aby:

$$\widehat{f} = \arg\inf_{f\in F} \widehat{\mathrm{er}}_{\mathbf{z}}(f). \tag{7}$$

Takie postępowanie nazywane jest w literaturze regułą indukcyjną SAE (ang. *sample error minimization*) lub ERM (ang. *empirical error minimization*)[6]. Innymi słowy na sam algorytm uczący możemy patrzeć jak na następujące odwzorowanie

$$L\colon \bigcup_{m=1}^{\infty}(\mathbf{X}\times\mathbf{Y})^m \to F, \tag{8}$$

które dla danej próby $\mathbf{z}$ wskazuje określoną hipotezę $L(\mathbf{z}) = \widehat{f}$. Jawne rozróżnienie pomiędzy zbiorem funkcji a algorytmem uczącym jest przydatne. Możemy pomyśleć np. o sieci neuronowej, gdzie funkcjami w zbiorze F są kombinacje lub złożenia sigmoid, i uczyć tę sieć różnymi algorytmami L: klasycznym *back-propagation*, algorytmem *RPROP*, metodą największej wiarygodności itd. Inny przykład — zbiór funkcji mogą stanowić wielomiany, które można uczyć metodą najmniejszych kwadratów: bez regularyzacji na współczynniki, z regularyzacją $\mathcal{L}_2$, z regularyzacją $\mathcal{L}_1$, itp.

Jak można zauważyć, błąd na próbie jest w zapisie pokrewny do błędu prawdziwego. Odpowiednie całki zastąpiono sumami. Jednakże, jak wiadomo, nie należy sądzić, że dla wybranej funkcji $\widehat{f}$ wartość jej błędu na próbie (zdolność do odtwarzania) to dyskretny odpowiednik lub oszacowanie dla jej błędu prawdziwego (zdolność do uogólniania). W większości przypadków błędy na próbie są mniejsze niż

[4] Bardzo rzadko bywają używane inne funkcje niż kwadratowa funkcja błędu.

[5] Bywa też nazywany *ryzykiem empirycznym* (ang. *empirical risk*)

[6] O ile arg inf istnieje. Jeżeli F jest zbiorem zwartym, to wówczas funkcjonały $\widehat{\mathrm{er}}_{\mathbf{z}}$ i er_P zdefiniowane nad zbiorem zwartym osiągają swoje kresy na mocy twierdzenia Weierstrassa.

błędy prawdziwe, jako że są one (błędy na próbie) osiągnięte poprzez wybór funkcji dobrze dopasowanej do konkretnych danych uczących, a tym samym szumów obecnych w tych danych. Mówiąc inaczej, można łatwo wskazać funkcję, dla której błąd na próbie jest bliski zeru lub nawet zero, a jednocześnie funkcja ta słabo uogólnia tj. ma duży błąd prawdziwy.

Niech f^* oznacza najlepszą funkcję w zbiorze F, taką że:

$$f^* = \arg\inf_{f \in F} \mathrm{er}_P(f). \tag{9}$$

Oczywiście chcielibyśmy, żeby funkcja $\widehat{f}$ wybrana przez algorytm uczący miała błąd prawdziwy $\mathrm{er}_P(\widehat{f})$ jak najbliższy do $\mathrm{er}_P(f^*)$.

Oprócz rozważania zbioru F dobrze jest w pewnych kontekstach patrzeć równolegle na **zbiór funkcji błędu** (ang. *loss functions*). Chodzi tu o zbiór: $l_F = \{l_f \colon f \in F\}$, gdzie dla zadania klasyfikacji mamy funkcje $l_f(\mathbf{z}) = l_f((\mathbf{x}, y)) = [f(\mathbf{x}) \neq y]$ realizujące odwzorowanie zerojedynkowe $l_f \colon \mathbf{X} \times Y \to \{0, 1\}$, a dla zadania estymacji regresji mamy funkcje $l_f(\mathbf{z}) = (f(\mathbf{x}) - y)^2$ realizujące odwzorowanie $l_f \colon \mathbf{X} \times Y \to \mathbb{R}$. Pomiędzy zbiorami F i l_F istnieje odpowiedniość $1 : 1$. Warto dodatkowo zauważyć, że dla zadania klasyfikacji, niezależnie od liczby klas w problemie (tj. niezależnie od liczności przeciwdziedziny, do której odwzorowują funkcje $f \colon \mathbf{X} \to Y$) funkcje l_f są zawsze funkcjami zerojedynkowymi. Ten fakt ma znaczenie przy definicji wymiaru Vapnika-Chervonenkisa, o którym później.

Zbieżność jednostajna dla skończonych zbiorów funkcji zerojedynkowych

Jednym z podstawowych celów teorii SLT jest badanie tempa **jednostajnej zbieżności w prawdopodobieństwie** (ang. *uniform convergence in probability*) błędów na próbie do błędów prawdziwych, gdyby generować ciąg takich wyników wraz z podnoszeniem rozmiaru m próby uczącej. Jednostajność oznacza, że interesuje nas najbardziej pesymistyczny przypadek, który może mieć miejsce tj. pytamy o prawdopodobieństwo takiego zdarzenia, że $\sup_{f \in F} |\mathrm{er}_P(f) - \widehat{\mathrm{er}}_{\mathbf{z}}(f)| > \epsilon$.

Jednym z narzędzi przydatnych do twierdzeń o jednostajnej zbieżności jest nierówność Chernoffa. Opisuje ona związek pomiędzy prawdopodobieństwem p pewnego zdarzenia, a jego częstością ν_m zaobserwowaną na próbie o rozmiarze m:

$$P_m(|p - \nu_m| > \epsilon) \leqslant 2\exp(-2\epsilon^2 m), \tag{10}$$

gdzie prawdopodobieństwo P_m jest wyliczane względem przestrzeni wszystkich prób o rozmiarze m. Jak widać prawdopodobieństwo odchyłki większej niż ϵ maleje w tempie wykładniczym ze względu na rozmiar próby. Istnieją też wersje jednostronne nierówności Chernoffa: $P_m(p - \nu_m > \epsilon) \leqslant \exp(-2\epsilon^2 m)$ i $P_m(\nu_m - p > \epsilon) \leqslant \exp(-2\epsilon^2 m)$.

Rozważmy najprostszy przypadek *skończonego* zbioru funkcji $F = \{f_1, \ldots, f_N\}$ użytego do uczenia. Znany jest następujący elementarny rezultat (Vapnik i Chervonenkis, 1971; Vapnik, 1998) o jednostajnej zbieżności:

$$P_m\left(\sup_{f \in F} |\mathrm{er}_P(f) - \widehat{\mathrm{er}}_{\mathbf{z}}(f)| > \epsilon\right) \leqslant \sum_{k=1}^{N} P_m\left(|\mathrm{er}_P(f_k) - \widehat{\mathrm{er}}_{\mathbf{z}}(f_k)| > \epsilon\right) \leqslant N \cdot 2\exp(-2\epsilon^2 m), \tag{11}$$

gdzie ostatnie przejście wynika z faktu, że dla każdej ustalonej funkcji f_k zachodzi nierówność Chernoffa[7]. Przypisując do prawej strony nierówności (11) pewne małe prawdopodobieństwo δ i rozwiązując

[7]Która stosuje się, ponieważ dla klasyfikacji wielkości er_P i $\widehat{\mathrm{er}}_{\mathbf{z}}$ oznaczają odpowiednio prawdopodobieństwo i częstość błędnego sklasyfikowania.

ze względu na ϵ, powyższy rezultat można równoważnie wyrazić w formie **ograniczenia na błąd prawdziwy**[8]:

$$\mathrm{er}_P(f_k) \leqslant \widehat{\mathrm{er}}_{\mathbf{z}}(f_k) + \sqrt{\frac{\ln N - \ln \delta}{2m}}, \tag{12}$$

które zachodzi z prawdopodobieństwem przynajmniej $1 - \delta$ dla *każdej* funkcji f_k w zbiorze F. W szczególności zachodzi też więc dla $\widehat{f}$. Idąc dalej można łatwo pokazać[9], że z prawdopodobieństwem przynajmniej $1 - 2\delta$:

$$\mathrm{er}_P(\widehat{f}) - \mathrm{er}_P(f^*) \leqslant \sqrt{\frac{\ln N - \ln \delta}{2m}} + \sqrt{\frac{-\ln \delta}{2m}}, \tag{13}$$

co stanowi ograniczenie na różnicę pomiędzy błędem prawdziwym wybranej funkcji $\widehat{f}$ a błędem prawdziwym najlepszej możliwej funkcji f^* w przyjętym F. Należy przypomnieć, że oba te błędy prawdziwe są w praktyce nieznane, a mimo to — co ciekawe — podanie ograniczenia jest możliwe.

Złożoność próbkowa

Kolejnym ważnym pojęciem jest **złożoność próbkowa** $m_L(\epsilon, \delta)$ — jest to minimalny rozmiar próby wystarczający na uczenie algorytmem L z zadaną (ϵ, δ)-precyzją dla danego problemu. Ograniczenia na złożoność próbkową otrzymuje się bezpośrednio z ograniczeń w stylu nierówności (13)[10]. I tak dla uproszczonego przypadku skończonego zbioru funkcji złożoność próbkowa jest ograniczona następująco:

$$m_L(\epsilon, \delta) \leqslant \frac{1}{\epsilon^2} \left(\sqrt{\ln N - \ln(\delta/2)} + \sqrt{-\ln(\delta/2)} \right)^2. \tag{14}$$

W omówionym przypadku wielkością reprezentującą bogatość zbioru F była liczba N — liczba funkcji w zbiorze. Oczywiście nie jest to przypadek praktyczny, i tak naprawdę w praktyce interesuje nas uczenie w oparciu o nieskończone zbiory funkcji (continuum funkcji). Dla tych zbiorów, ograniczenia na błąd prawdziwy i złożoność próbkową są budowane w oparciu o inne pojęcia bogatości, pojemności F. Mówiąc skrótowo należy zastąpić pojawiający się $\ln N$ pewnym odpowiednikiem właściwym dla nieskończonego zbioru F. I tak dla nieskończonych zbiorów funkcji zerojedynkowych (klasyfikacja) jest to zwykle logarytm z tzw. *funkcji wzrostu*, a dla nieskończonych zbiorów funkcji rzeczywistych (estymacja regresji) jest to zwykle logarytm z *liczby pokryciowej*.

Zbieżność jednostajna dla nieskończonych zbiorów funkcji zerojedynkowych

Niech F oznacza nieskończony zbiór funkcji. Dla ustalonej próby $\mathbf{z}_1, \ldots, \mathbf{z}_m$ niech $(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m}$ oznacza zbiór funkcji błędu rozróżnialnych nad tą próbą (lub inaczej: zbiór funkcji odciętych do próby), tj.:

$$(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = \Big\{ \big(l_f(\mathbf{z}_1), \ldots, l_f(\mathbf{z}_m) \big) \colon\ f \in F \Big\}. \tag{15}$$

Oczywiście $\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} \leqslant 2^m$. Ważnym pojęciem w tym kontekście jest **roztrzaskiwanie** (ang. *shattering*). Mówimy, że zbiór funkcji zerojedynkowych *roztrzaskuje* próbę $\mathbf{z}_1, \ldots, \mathbf{z}_m$, jeżeli w tym zbiorze istnieje 2^m funkcji rozróżnialnych nad próbą[11]. Idąc dalej, za pomocą roztrzaskiwania definiowany jest **wymiar Vapnika-Chervonenkisa** (Vapnik i Chervonenkis, 1971).

[8] Użyto jednostronnej wersji nierówności Chernoffa, ponieważ interesuje nas ograniczenie z góry.

[9] Wystarczy wykorzystać dwa fakty: (1) z definicji $\widehat{f}$ mamy $\widehat{\mathrm{er}}_{\mathbf{z}}(f^*) \geqslant \widehat{\mathrm{er}}_{\mathbf{z}}(\widehat{f})$, oraz (2) dla f^* zachodzi bezpośrednio nierówność Chernoffa.

[10] Należy przypisać ϵ do lewej strony nierówności, uzgodnić prawdopodobieństwo $\delta := \delta/2$ i rozwiązać ze względu na m.

[11] Inaczej mówiąc, można zrealizować wszystkie dychotomie próby za pomocą funkcji z tego zbioru.

Definicja 1 *Mówimy, że zbiór l_F funkcji zerojedynkowych ma wymiar Vapnika-Chervonenkisa równy h, VC-dim $(l_F) = h$, jeżeli istnieje próba o rozmiarze h roztrzaskiwana przez l_F i nie istnieje żadna taka próba o rozmiarze $h+1$. Jeżeli dla każdego $h > 0$ istnieje pewna próba roztrzaskiwana, to VC-dim $(l_F) = \infty$.*

Znane są pewne zbiory funkcji, dla których dokładna wartość wymiaru VC została ustalona poprzez odpowiedni dowód kombinatoryczny. Oto niektóre przykłady. Dla wielomianów zdefiniowanych nad $\mathbb{R}^d$ stopnia co najwyżej n, wymiar VC wynosi $\binom{n+d}{d}$, patrz np. (Anthony i Bartlett, 2009). Dla płaszczyzn w $\mathbb{R}^d$, które mogą być bazami w sieciach perceptronowych, wymiar VC wynosi $d+1$ (Vapnik, 1998). Dla kostek w $\mathbb{R}^d$ wymiar VC wynosi $2d$ (Cherkassky i Mulier, 1998). Dla kul w $\mathbb{R}^d$, które mogą być bazami dla radialnych sieci neuronowych, wymiar VC wynosi $d+1$ (Cherkassky i Mulier, 1998). Jeżeli chodzi o liniowe kombinacje baz, to wymiar VC można zwykle ograniczyć z góry przez iloczyn liczby baz i wymiaru VC pojedynczej bazy (Anthony i Bartlett, 2009, str. 154), ale to stwierdzenie wymaga zwykle ostrożnej analizy. Warto dodać jednocześnie, że istnieją zbiory funkcji, dla których nie udało się jeszcze określić (dowieść) wymiaru VC, a mimo to zbiory te są wykorzystywane w uczeniu[12].

Alternatywnie, wymiar VC można definiować w oparciu o **funkcję wzrostu** (ang. *growth function*), która podaje największą liczbę rozróżnialnych funkcji dla danego rozmiaru próby: $G^F(m) = \sup_{\mathbf{z}\in(\mathbf{X}\times Y)^m} \#(l_F)_{|\mathbf{z}}$. Wymiar VC, to największy argument funkcji wzrostu, powyżej którego przestaje ona narastać wykładniczo. Jednym z istotnych rezultatów w tym kontekście jest lemat Sauera (Sauer, 1972; Steel, 1978; Chari, Rohatgi i Srinivasan, 1994), który mówi, że jeżeli VC-dim$(l_F) = h$, to:

$$G^F(m) \leqslant \sum_{i=0}^{h} \binom{m}{h} < \left(\frac{me}{h}\right)^h. \tag{16}$$

Poniższe twierdzenie o jednostajnej zbieżności (z wykorzystaniem lematu Sauera) sformułowali oryginalnie Vapnik i Chervonenkis (1971), patrz także (Anthony i Bartlett, 2009).

Twierdzenie 1 *Niech F będzie nieskończonym zbiorem funkcji zerojedynkowych o funkcji wzrostu $G^F(m)$ i VC-dim$(F) = h$. Wówczas:*

$$P_m\left(\sup_{f\in F} |er_P(f) - \widehat{er}_{\mathbf{z}}(f)| \geqslant \epsilon\right) \leqslant 4G^F(2m)\exp\left(-m\epsilon^2/8\right) \tag{17}$$

$$\leqslant 4\exp\left(h\left(1+\ln\frac{2m}{h}\right) - m\epsilon^2/8\right). \tag{18}$$

Analogicznie do wzoru (12), powyższy rezultat można zapisać równoważnie w formie ograniczenia na błąd prawdziwy — z prawdopodobieństwem przynajmniej $1-\delta$ dla każdej funkcji $f \in F$ mamy

$$\mathrm{er}_P(f) \leqslant \widehat{\mathrm{er}}_{\mathbf{z}}(f) + \sqrt{\frac{h(1+\ln(2m/h)) - \ln(\delta/4)}{m/8}}. \tag{19}$$

Funkcje rzeczywiste w uczeniu, pokrycia, liczby pokryciowe

W przypadku nieskończonych zbiorów funkcji zerojedynkowych wykorzystywany był fakt, że dla każdej ustalonej próby zbiór funkcji odciętych do próby $(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m}$ stawał się zbiorem skończonym. W uczeniu za pomocą funkcji rzeczywistych zbiór ten jest niestety nadal nieskończony — dziedzinę stanowi skończona liczba punktów, ale na przeciwdziedzinie nadal mamy continuum wartości. To uniemożliwia

[12] M.in.: klasyfikatory probabilistyczne, klasyfikatory oparte na łańcuchach Markowowskich, krzywe sklejane MARS, niektóre drzewa decyzyjne z procedurą przycinania.

zliczanie. Potrzebnym zabiegiem staje się zastosowanie pojęcia **pokrycia** (ang. *cover*), co pozwala na redukcję zbioru nieskończonego do skończonego i w efekcie na zliczanie.

W ogólności mówimy, że zbiór U jest ϵ-pokryciem zbioru W zawartego w przestrzeni metrycznej, jeżeli dla każdego $w \in W$ istnieje element $u \in U$, taki że: $d(u, w) < \epsilon$ (gdzie d oznacza przyjętą metrykę). W problemach uczenia interesuje nas pokrywanie zbioru $(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m}$, który stanowi pewne **zamazanie** zbioru $\mathbb{R}^m$. Jeżeli funkcje l_f są ograniczone, to mówimy o zamazaniu pewnej kostki zawartej w $\mathbb{R}^m$. Należy dodać, że istnieją twierdzenia, które pozwalają wyrazić pokrycie zbioru $(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m}$ w terminach pokrycia zbioru $F_{|\mathbf{x}_1,\ldots,\mathbf{x}_m}$. Jest to udogodnienie, ponieważ zbiorem F zajmujemy się bezpośrednio.

Definicja 2 ***Liczbą pokryciową*** $\mathcal{N}(\epsilon, F_{|\mathbf{x}_1,\ldots,\mathbf{x}_m}, d)$ *nazywamy rozmiar minimalnego ϵ-pokrycia zbioru $F_{|\mathbf{x}_1,\ldots,\mathbf{x}_m}$ w metryce d.*

Definicja 3 ***Jednostajną liczbą pokryciową*** $\mathcal{N}_d(\epsilon, F, m)$ *nazywamy maksymalną spośród liczb $\mathcal{N}(\epsilon, F_{|\mathbf{x}_1,\ldots,\mathbf{x}_m}, d)$ biorąc pod uwagę wszystkie możliwe próby o danym rozmiarze m:*

$$\mathcal{N}_d(\epsilon, F, m) = \max\{\mathcal{N}(\epsilon, F_{|\mathbf{x}_1,\ldots,\mathbf{x}_m}, d) : \mathbf{x} \in \mathbf{X}^m\}. \tag{20}$$

Należy zaznaczyć, że dla liczb pokryciowych używa się w ogólności metryk w postaci $d_q(u, w) = (1/m \sum_i |u_i - v_i|^q)^{1/q}$. Z uwagi na czynnik $1/m$ zachodzi relacja: $\mathcal{N}_1(\cdot) \leqslant \mathcal{N}_2(\cdot) \leqslant \mathcal{N}_\infty(\cdot)$.

Istnieją następujące ważne twierdzenia o zbieżności jednostajnej wykorzystujące liczby pokryciowe.

Twierdzenie 2 *(Anthony i Bartlett, 2009, twierdzenie 10.1) Niech F oznacza zbiór funkcji f: $\mathbf{X} \to [0, 1]$, a P łączny rozkład prawdopodobieństwa zdefiniowany nad $\mathbf{X} \times \{0, 1\}$. Niech: $0 < \epsilon < 1$, $\gamma > 0$. Wtedy*

$$P_m\left(\sup_{f \in F} er_P(f) - \widehat{er}^{\gamma}_{\mathbf{z}}(f) \geqslant \epsilon\right) \leqslant 2\mathcal{N}_\infty(\gamma/2, F, 2m) \exp(-\epsilon^2 m/8). \tag{21}$$

Twierdzenie 3 *(Anthony i Bartlett, 2009, twierdzenie 17.1) Niech F oznacza zbiór funkcji f: $\mathbf{X} \to [0, 1]$, a P łączny rozkład prawdopodobieństwa zdefiniowany nad $\mathbf{X} \times [0, 1]$. Niech: $0 < \epsilon < 1$. Wtedy*

$$P_m\left(\sup_{f \in F} |er_P(f) - \widehat{er}_{\mathbf{z}}(f)| \geqslant \epsilon\right) \leqslant 4\mathcal{N}_1(\epsilon/16, F, 2m) \exp(-\epsilon^2 m/32). \tag{22}$$

Twierdzenie 2 jest sformułowane dla zadania klasyfikacji postawionego jako **klasyfikacja z marginesem** γ i wykorzystuje liczbę pokryciową w metryce d_∞. Margines reprezentuje odległość od progowej granicy decyzji, przy czym odległość ta jest liczona na osi wartości funkcji f, tzn.: $\text{margin}(f(\mathbf{x}), y) = f(\mathbf{x}) - \frac{1}{2}$ dla $y = 1$ oraz $\text{margin}(f(\mathbf{x}), y) = \frac{1}{2} - f(\mathbf{x})$ dla $y = 0$. Intuicyjnie: im większy margines, tym pewniejsze zaklasyfikowanie (Bartlett, 1998). W twierdzeniu pojawia się $\widehat{\text{er}}^{\gamma}_{\mathbf{z}}$, co oznacza częstość marginesu mniejszego niż γ na próbie, tj.: $\widehat{\text{er}}^{\gamma}_{\mathbf{z}}(f) = \frac{1}{m}\sum_{i=1}^{m}[\text{margin}\,(f(\mathbf{x}_i), y_i) < \gamma]$. Marginesu w powyższym rozumieniu nie należy mylić z pojęciem *marginesu separacji* w maszynach SVM. Tam margines liczony jest w przestrzeni $\mathbf{X}$ a nie na osi wartości funkcji f. Twierdzenie 3 jest sformułowane dla zadania estymacji regresji i wykorzystuje liczbę pokryciową w metryce d_1.

Następujące rezultaty są trzema przykładowymi ograniczeniami na liczby pokryciowe. Dwa pierwsze z nich wykorzystują *pseudowymiar*[13] (ang. *pseudodimension*) jako pojęcie pojemności zbioru funkcji.

[13]Pseudowymiar jest tym dla funkcji rzeczywistych, czym wymiar VC dla funkcji zerojedynkowych — w praktyce można utożsamiać te dwa pojęcia, a ich liczbowa wartość jest taka sama po zaokrągleniu rzeczywistych wartości f do $\{0, 1\}$.

Ostatni rezultat (twierdzenie 6) wyraża liczbę pokryciową w terminach uczenia z **regularyzacją** dla funkcji liniowych ze względu na parametry. Regularyzacja powoduje, że oprócz błędu na próbie minimalizujemy także normę parametrów $\|w\|$ w pewnej metryce.

Twierdzenie 4 *(Haussler i Long, 1995) Niech F oznacza zbiór funkcji rzeczywistych $f\colon \mathbf{X} \to [0,1]$ o pseudowymiarze równym h. Wtedy:*

$$\mathcal{N}_\infty(\epsilon, F, m) \leqslant \sum_{i=0}^{h} \binom{m}{i} \lfloor 1/\epsilon \rfloor^i, \tag{23}$$

co z kolei jest mniejsze niż $\left(\frac{me}{\epsilon h}\right)^h$ dla $m \geqslant h$.

Twierdzenie 5 *(Haussler, 1995) Niech F oznacza zbiór funkcji rzeczywistych $f\colon \mathbf{X} \to [0,1]$ o pseudowymiarze równym h. Wtedy:*

$$\mathcal{N}_1(\epsilon, F, m) \leqslant e(h+1)\left(\frac{2e}{\epsilon}\right)^h. \tag{24}$$

Twierdzenie 6 *(Zhang, 2002) Niech F będzie zbiorem funkcji liniowych postaci: $f(\mathbf{x}) = \sum_{j=1}^{d} w_j x_j$, i niech algorytm uczący $\mathcal{L}_q$-regularyzuje wagi, tj. mamy, że $\|w\|_q \leqslant a$. Dla ustalonego q, zbiór danych jest znormalizowany następująco: $\|\mathbf{x}_i\|_p \leqslant b$, $i = 1, \ldots, m$, gdzie $1/p + 1/q = 1$ (normy sprzężone) oraz $2 \leqslant p \leqslant \infty$. Wtedy:*

$$\mathcal{N}_2(\epsilon, F, m) \leqslant (2d+1)^{\lceil a^2 b^2/\epsilon^2 \rceil}. \tag{25}$$

Twierdzenie 6 to bardzo atrakcyjny wynik. Po przełożeniu na złożoność próbkową mówi on, że przy zastosowaniu regularyzacji, do uczenia się z precyzją (ϵ, δ) wystarcza próba o rozmiarze proporcjonalnym tylko do logarytmu z liczby atrybutów (a nie skalująca się liniowo wraz z liczbą atrybutów). Przypomnijmy, że w wyrażeniu na złożoność próbkową pojawi się wyraz $\ln \mathcal{N}_1(\cdot)$ oraz że $\mathcal{N}_1(\cdot) \leqslant \mathcal{N}_2(\cdot)$. I tak po zlogarytmowaniu (25) otrzymamy $\lceil a^2 b^2/\epsilon^2 \rceil \ln(2d+1) = O(\ln d)$, zaś po zlogarytmowaniu (24) otrzymamy $(d+1)\ln(2e/\epsilon) + \ln(d+2) + 1 = O(d)$, wiedząc, że wymiar VC wynosi $h = d + 1$ dla funkcji liniowo zależnych od parametrów.

4b.2 Rezultaty habilitanta

Rezultaty habilitanta dotyczą w dużej mierze różnych technik wyboru złożoności modelu. Jak wiadomo oprócz technik opartych na wielokrotnym testowaniu (np. n-krotna krzyżowa walidacja, bootstrap) można stosować techniki oparte na ograniczeniach na błąd prawdziwy, zobacz np. nierówności (12), (19), (21), (22). Idea tego drugiego podejścia została zapoczątkowana przez Vapnika i znana jest pod nazwą **SRM** (ang. *Structural Risk Minimization*) (Vapnik, 1995). Ogólny schemat jest następujący. Przeprowadzany jest ciąg eksperymentów, w którym stopniowo podnosimy złożoność (pojemność) zbioru funkcji i zamiast patrzeć na średni błąd testowania, patrzymy na ograniczenia na błąd prawdziwy. Oczywiście w takim ciągu otrzymujemy coraz mniejsze błędy uczące $\widehat{\mathrm{er}}_{\mathbf{z}}$ — pierwszy składnik w ograniczeniach, ale jednocześnie powiększamy drugi składnik związany ze złożonością modelu i coraz gorszą zdolnością do uogólniania. Ostatecznie wybieramy złożoność odpowiadającą *punktowi minimum* tych ograniczeń.

Związek pomiędzy krzyżową walidacją a ograniczeniami Vapnika na błąd prawdziwy

W pracy (Klęsk, 2011) habilitant podaje twierdzenia pokazujące probabilistyczny związek pomiędzy ograniczeniami Vapnika a wynikami n-krotnej krzyżowej walidacji. Jak wiadomo przeprowadzenie krzyżowej walidacji jest około n razy kosztowniejsze obliczeniowo niż procedura SRM. W pracy przyjęto następujące podejście i cele.

(a) Nie koncentrujemy się na tym, na ile dobrze zmierzony wynik krzyżowej walidacji (średni błąd testowania) przybliża błąd prawdziwy. Zamiast tego, **chcemy móc wypowiadać się o wyniku krzyżowej walidacji** dla danych warunków eksperymentu **nie wykonując jej**.

(b) Chcemy z dokładnością probabilistyczną **podać, jakiej różnicy ϵ można oczekiwać pomiędzy *znaną* (wyliczoną) wartością ograniczeń Vapnika a *nieznanym* wynikiem krzyżowej walidacji**.

(c) W konsekwencji, chcemy wyznaczyć potrzebny rozmiar próby, taki żeby ϵ był wystarczająco mały; dzięki czemu wskazana przez procedurę SRM optymalna złożoność jest akceptowalna w tym sensie, że krzyżowa walidacja (gdyby została wykonana) wskazałaby prawdopodobnie tę samą złożoność.

W pracy rozważany jest wariant *non-stratified* krzyżowej walidacji (Kohavi, 1995) — wygodny na potrzeby twierdzeń. Wariant ten wprowadza niezależność pomiędzy poszczególnymi iteracjami walidacji[14] i może być wyrażony algorytmem:

1 Dla $k = 1, 2, \ldots, n$ powtarzaj:

  1.1 Przepermutuj zbiór danych.

  1.2 Podziel dane w ustalonym punkcie $\frac{n-1}{n}m$ na: *zbiór uczący* $\{\mathbf{z}'_1, \mathbf{z}'_2, \ldots, \mathbf{z}'_{m'}\}$ i *zbiór testowy* $\{\mathbf{z}''_1, \mathbf{z}''_2, \ldots, \mathbf{z}''_{m''}\}$, gdzie $m' = \frac{n-1}{n}m$ and $m'' = \frac{1}{n}m$ (z zaokrągleniem całkowitym).

  1.3 Znajdź funkcję $\widehat{f'_k}$ minimalizującą błąd uczący dla aktualnej k-tej iteracji, t.ż.:
  $\widehat{f'_k} = \arg\inf_{f \in F} \frac{1}{m'} \sum_{i=1}^{m'} l_f(\mathbf{z}'_i)$.

  1.4 Oblicz błąd testowy dla $\widehat{f'_k}$:
  $\widehat{\mathrm{er}}_{\mathbf{z}''}(\widehat{f'_k}) = \frac{1}{m''} \sum_{i=1}^{m''} l_{\widehat{f'_k}}(\mathbf{z}''_i)$.

2 Oblicz wynik krzyżowej walidacji jako średnią z wyników testów:

$$C = \frac{1}{n} \sum_{k=1}^{n} \widehat{\mathrm{er}}_{\mathbf{z}''}(\widehat{f'_k}). \tag{26}$$

Poniższe twierdzenia pokazują związek wielkości C z ograniczeniem Vapnika na błąd prawdziwy (nazwijmy je V) dla uproszczonego przypadku skończonego zbioru funkcji $F = \{f_1, \ldots, f_N\}$. Ograniczenie V ma postać:

$$V = \widehat{\mathrm{er}}_{\mathbf{z}}(\widehat{f}) + \sqrt{\frac{\ln N - \ln \delta}{2m}}, \tag{27}$$

[14] Podział na zbiór testowy i uczący jest dokonywany niezależnie w każdej iteracji bez „oglądania się" na pozostałe iteracje. W efekcie zbiory uczące nie muszą być parami rozłączne (podobnie zbiory testowe), a dbamy tylko o zachowanie proporcji rozmiarów zbiorów uczących do testowych.

gdzie $\widehat{f} = \arg\inf_{f\in F} \widehat{\mathrm{er}}_{\mathbf{z}}(f)$. Należy dodać, że podane twierdzenia można natychmiastowo uogólnić na przypadek nieskończonego zbioru funkcji, zastępując odpowiednio wyrażenie pod pierwiastkiem, patrz np. (19) — w szczególności zastępując $\ln N$ przez odpowiednie wyrażenie związane z pojemnością nieskończonego zbioru F.

Twierdzenie 7 *(Klęsk, 2011, twierdzenie 1) Niech $F = \{f_1, \ldots, f_N\}$ oznacza skończony zbiór funkcji zerojedynkowych. Wtedy, dla każdego $\delta > 0$, istnieje mała liczba*

$$\alpha(\delta, n) = \delta - \sum_{k=1}^{n} \binom{n}{k} (-1)^k (2\delta)^k, \tag{28}$$

oraz liczba

$$\epsilon(\delta, m, N, n) = \left(2\sqrt{\frac{n}{n-1}} + 1\right)\sqrt{\frac{\ln N - \ln\delta}{2m}} + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}}\right)\sqrt{\frac{-\ln\delta}{2m}}, \tag{29}$$

taka że:

$$P\left(|V - C| \leq \epsilon(\delta, m, N, n)\right) \geq 1 - \alpha(\delta, n). \tag{30}$$

Uwaga 1 Wartość $\alpha(\delta, n)$ jest monotoniczna z δ. To znaczy im mniejsze wybierzemy δ, tym mniejsze jest również $\alpha(\delta, n)$. Stąd, minimalna wartość prawodpodobieństwa $1 - \alpha(\delta, n)$ jest wystarczająco duża.

$$\lim_{\delta\to 0^+}\left(\delta - \sum_{k=1}^{n}\binom{n}{k}(-1)^k(2\delta)^k\right) = \lim_{\delta\to 0^+}\left(\delta + 1 - \sum_{k=0}^{n}\binom{n}{k}(-1)^k(2\delta)^k\right) = \lim_{\delta\to 0^+}(\delta + 1 - (1-2\delta)^n) = 0. \tag{31}$$

Uwaga 2 Dla ustalonych δ, N, n, wartość $\epsilon(\delta, m, N, n)$ zbiega do zera wraz ze wzrostem rozmiaru próby.

Poniższe twierdzenia 8, 9 stanowią odpowiednio górne i dolne ograniczenie na C, patrz Rys. 1a. Dowiedzenie ich natychmiastowo dowodzi głównego twierdzenia 7. Dowody zawarte w (Klęsk, 2011) opierają się na wielokrotnym złożeniu nierówności Chernoffa.

Twierdzenie 8 *(Klęsk, 2011, twierdzenie 2) Z prawdopodobieństwem $1 - \alpha(\delta, n)$ lub większym prawdziwa jest nierówność:*

$$C - V \leq \left(\sqrt{\frac{n}{n-1}} - 1\right)\sqrt{\frac{\ln N - \ln\delta}{2m}} + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}}\right)\sqrt{\frac{-\ln\delta}{2m}}. \tag{32}$$

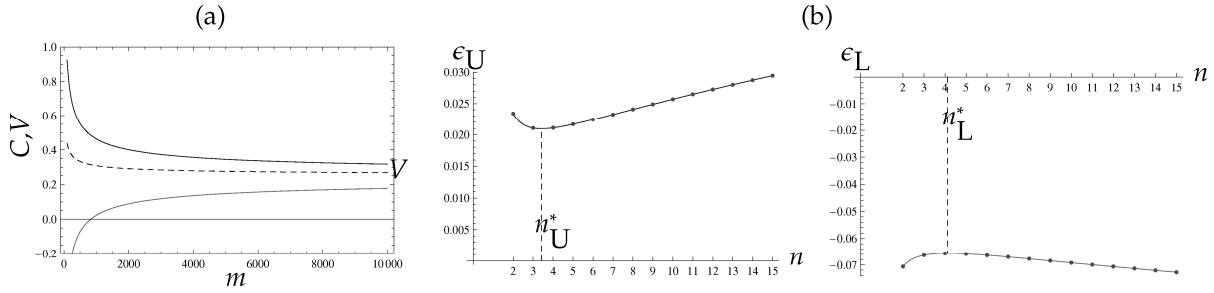
Twierdzenie 9 *(Klęsk, 2011, twierdzenie 3) Z prawdopodobieństwem $1 - \alpha(\delta, n)$ lub większym prawdziwa jest nierówność:*

$$V - C \leq \left(2\sqrt{\frac{n}{n-1}} + 1\right)\sqrt{\frac{\ln N - \ln\delta}{2m}} + \sqrt{n}\sqrt{\frac{-\ln\delta}{2m}}. \tag{33}$$

W konsekwencji, zadanie górnego i dolnego ograniczenia epsilonowego $\epsilon^*_{\mathrm{U}}, \epsilon^*_{\mathrm{L}}$ powoduje, że rozmiary próby potrzebne do tego, aby wynik krzyżowej walidacji C był niegorszy niż $V + \epsilon^*_{\mathrm{U}}$ i nielepszy niż $V - \epsilon^*_{\mathrm{L}}$ wynoszą odpowiednio:

$$m(\epsilon^*_{\mathrm{U}}, \delta) \geqslant \frac{1}{2{\epsilon^*_{\mathrm{U}}}^2}\left(\left(\sqrt{\frac{n}{n-1}} - 1\right)\sqrt{\ln N - \ln\delta} + \left(\sqrt{n} + \sqrt{\frac{n}{n-1}}\right)\sqrt{-\ln\delta}\right)^2 \tag{34}$$

$$m(\epsilon^*_{\mathrm{L}}, \delta) \geqslant \frac{1}{2{\epsilon^*_{\mathrm{L}}}^2}\left(\left(2\sqrt{\frac{n}{n-1}} + 1\right)\sqrt{\ln N - \ln\delta} + \sqrt{n}\sqrt{-\ln\delta}\right)^2. \tag{35}$$



Rysunek 1: (a) Ilustracja górnego i dolnego ograniczenia na wynik krzyżowej walidacji zawężającego się wraz ze wzrostem rozmiaru próby m. Pozostałe stałe to: $\delta = 0.01 \Rightarrow 1 - \alpha(\delta) \approx 0.93$, $N = 100$, $n = 3$. (b) Najlepsze wartości liczby n iteracji walidacyjnych dla zawężenia ograniczeń.

Poniższa obserwacja, jest szczególnym następstwem twierdzeń.

Konsekwencja 1 Dla krzyżowej walidacji *leave-one-out*, gdzie $n = m$, oba ograniczenia górne i dolne poluzowują się do stałej rzędu $O(\sqrt{-1/2 \ln \delta})$.

Dodatkowo, można też postawić takie ciekawe pytanie: dla jakiego wyboru liczby n iteracji krzyżowej walidacyji, wyznaczone ograniczenia na wynik C stają się najbardziej ciasne przy ustalonych δ, m, N? Traktując chwilowo n jako zmienną ciągłą, żądamy aby $\partial \epsilon_U(\delta, m, N, n)/\partial n = 0$ oraz $\partial \epsilon_L(\delta, m, N, n)/\partial n = 0$ i otrzymujemy optymalne wartości n (co zilustrowano na Rys. 1b):

$$n^*_U = 1 + \left(\frac{\sqrt{\ln N - \ln \delta} + \sqrt{-\ln \delta}}{\sqrt{-\ln \delta}} \right)^{2/3}, \tag{36}$$

$$n^*_L = 1 + \left(\frac{2\sqrt{\ln N - \ln \delta}}{\sqrt{-\ln \delta}} \right)^{2/3}. \tag{37}$$

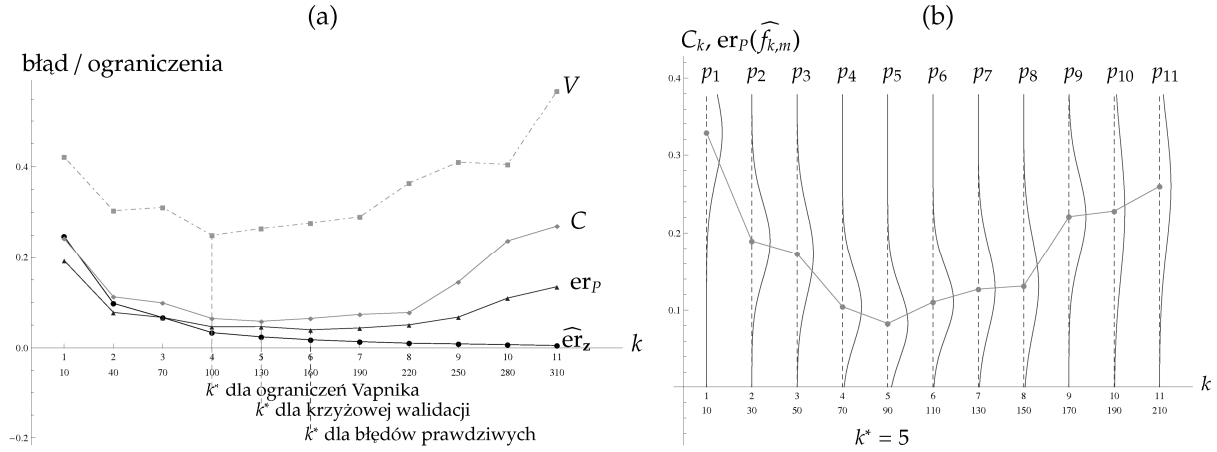
Proszę zwrócić tu uwagę na brak zależności tych wartości — co ciekawe — od rozmiaru próby.

W pracy (Klęsk, 2011) przedstawiono także eksperymenty, w których obserwowano, jak często wynik C krzyżowej walidacji wpadał w przedział $[V - \epsilon^*_L, V + \epsilon^*_U]$ dla zadanych warunków doświadczenia. Wyniki eksperymentów potwierdzały wyniki teoretyczne.

Prawdopodobieństwa rozbieżności pomiędzy punktami minimum procedur wyboru złożoności i nieznanym punktem minimum błędu prawdziwego

Stosując pewną procedurę wyboru złożoności modelu mamy do czynienia z problemem możliwej **rozbieżności** pomiędzy punktem minimum wskazanym empirycznie przez tę procedurę oraz faktycznym minimum błędu prawdziwego, która to wielkość nie jest znana. Taką sytuację zilustrowano na Rys. 2a. Problem jest oczywisty i intuicyjny, ale w praktyce zaniedbywany. W pracy (Klęsk, 2010*b*) habilitant stawia ten problem w ilościowej i nowatorskiej postaci, a mianowicie: jako **problem obliczenia prawdopodobieństw zdarzenia, że taka rozbieżność zachodzi (lub nie zachodzi)**. Rozwiązanie problemu daje dodatkową (ilościową) wiedzę o niepewności eksperymentu.

Główne twierdzenia w (Klęsk, 2010*b*) są w oryginalnej postaci sformułowane dla zadania estymacji regresji i wyrażone w terminach wszystkich istotnych stałych: rozmiaru próby, liczby iteracji krzyżowej walidacji, pojemności zbioru funkcji, ograniczoności funkcji w tym zbiorze.



Rysunek 2: (a) Przykładowa ilustracja rozbieżności pomiędzy wskazaniami k^* punktów minimum: krzyżowej walidacji, ograniczeń Vapnika i błędu prawdziwego. Na osi argumentów położone są pozycje $k \in \{1, 2, \ldots, 11\}$ w strukturze sterujące złożonością (a pod nimi odpowiadająca im liczba termów w postaci funkcyjnej). Krzywe przedstawiają: błędy na próbie $\widehat{\mathrm{er}}_{\mathbf{z}}$, błędy prawdziwe er_P, wyniki krzyżowej walidacji C, wyniki ograniczeń Vapnika V. (b) Przykładowe wskazanie $k^* = 5$ optymalnej złożoności uzyskane krzyżową walidacją. W formie pionowej narysowano przybliżenia gęstości rozkładów prawdopodobieństwa na nieznany błąd prawdziwy.

Rozważmy następujący scenariusz. Zgodnie ze schematem SRM przebiegamy strukturę $F_1 \subset F_2 \subset \cdots \subset F_K$ zagnieżdżonych podzbiorów funkcji o wzrastającej złożoności. Dla każdego podzbioru F_k wykonujemy n-krotną krzyżową walidację *non-stratified*. Otrzymujemy jej wynik C_k. Przypomnijmy, że C_k stanowi oszacowanie *średniej z błędów prawdziwych*

$$C_k \sim \frac{1}{n} \sum_{j=1}^{n} \mathrm{er}_P(\widehat{f}_{k,m',j}) \tag{38}$$

wziętej po n funkcjach $\widehat{f}_{k,m',j}$ wskazanych przez algorytm uczący w poszczególnych iteracjach walidacyjnych dla $j = 1, \ldots, n$, używając każdorazowo próby uczącej $\mathbf{z}'$ o rozmiarze $m' = \frac{n-1}{n} m$. Kiedy procedura jest już zakończona dla całej struktury, mamy ciąg wyników $C_1, C_2, \ldots, C_K$ i wskazanie, że optymalna złożoność jest przypuszczalnie w punkcie k^*, t.ż.: $C_{k^*} = \min_{k \in \{1,\ldots,K\}} C_k$. Ostatecznie, możemy użyć do nauki całej próby o rozmiarze m (a nie tylko $\frac{n-1}{n} m$), i za pomocą algorytmu uczącego wybrać najlepszą funkcję pochodzącą ze zbioru F_{k^*} jako ostateczny model. Oznaczmy tę funkcję jako $\widehat{f}_{k^*,m}$. Można teraz postawić następujące dwa ważne pytania.

1. Jakie jest prawdopodobieństwo zdarzenia, że punkt k^* wskazany przez krzyżową walidację, jest naprawdę punktem minimum nieznanych błędów prawdziwych $\mathrm{er}_P(\widehat{f}_{k,m})$?

2. Z jakim prawdopodobieństwem prawdziwy punkt minimum wielkości $\mathrm{er}_P(\widehat{f}_{k,m})$ wpada w otoczenie wskazanego punktu k^* o pewnym promieniu Δ?

Innymi słowy, chcemy wiedzieć coś o wiarygodności wskazania k^* jako rzekomej optymalnej złożoności, lub przynajmniej chcielibyśmy wiedzieć, o jak dużo mogliśmy się pomylić. Poniżej przedstawiono dwa twierdzenia, które odpowiadają na postawione pytania, w taki sposób, że podają one minimalne (pesymistyczne) wartości interesujących nas prawdopodobieństw.

Twierdzenie 10 *(Klęsk, 2010*b*, twierdzenie 1) Niech $F_1 \subset F_2 \subset \cdots \subset F_K$ będzie strukturą zbiorów funkcji rzeczywistych i ograniczonych, t.ż. dla wszystkich $l_f \in l_{F_k}$ mamy $0 \leqslant l_f(\mathbf{z}) \leqslant B_k$. Niech każdy element F_k struktury ma skończoną pojemność N_k, t.j. skończoną liczbę N_k funkcji w zbiorze (dla przypadku zbiorów skończonych) lub skończony wymiar VC (dla zbiorów nieskończonych). Niech $C_1, C_2, \ldots, C_K$ będzie ciągiem wyników n-krotnej krzyżowej walidacji non-stratified otrzymanych dla tej struktury. Przypuśćmy, że minimum tych wyników jest osiągane w punkcie k^*:*

$$C_{k^*} = \min_{k\in\{1,\ldots,K\}} C_k. \tag{39}$$

Wtedy, minimalne prawdopodobieństwo, że wskazany punkt k^*, jest naprawdę punktem minimum nieznanych błędów prawdziwych $er_P(\widehat{f_{k,m}})$ może być obliczone następująco

$$P\Big(er_P(\widehat{f_{k^*,m}}) = \min_{k\in\{1,\ldots,K\}} er_P(\widehat{f_{k,m}})\Big) = \int_{-\infty}^{\infty}\Big(\prod_{k\neq k^*}\int_{r_{k^*}}^{\infty} p_k(r_k)dr_k\Big)p_{k^*}(r_{k^*})dr_{k^*}, \tag{40}$$

gdzie p_k są gęstościami rozkładów normalnych

$$p_k(r) \quad = \frac{1}{(1/\sqrt{n})\sqrt{{\sigma_{k1}}^2+{\sigma_{k2}}^2}\sqrt{2\pi}}\exp\Big(-\frac{(r-C_k)^2}{(2/n)({\sigma_{k1}}^2+{\sigma_{k2}}^2)}\Big) \tag{41}$$

o stałych

$$\begin{aligned}\sigma_{k1} &= \frac{B_k\sqrt{n}}{a_{1-\delta/2}}\sqrt{\frac{-\ln(\delta/2)}{2m}}\\ \sigma_{k2} &= \frac{B_k}{a_{1-\delta/2}}\Bigg(\sqrt{\frac{n}{n-1}}\sqrt{\frac{-\ln(\delta/6)}{2m}}+\Big(\sqrt{\frac{n}{n-1}}+1\Big)\sqrt{\frac{\ln N_k-\ln(\delta/6)}{2m}}\Bigg).\end{aligned} \tag{42}$$

$a_{1-\delta/2}$ oznacza kwantyl rzędu $1-\delta/2$ z rozkładu $N(0,1)$ dla dowolnego $\delta > 0$. Rozkłady normalne są przybliżeniami rozkładów na nieznane błędy prawdziwe, o jednostajnym[15] *błędzie przybliżenia rzędu $O\Big(\Big(1+\frac{1}{\sqrt{n-1}}+\sqrt{n}\Big)\frac{1}{\sqrt{m}}\Big)$.*

Twierdzenie 11 *(Klęsk, 2010*b*, twierdzenie 2) Minimalne prawdopodobieństwo, że prawdziwe minimum wielkości $er_P(\widehat{f_{k,m}})$ wpada w otoczenie $\{k\colon |k-k^*| \leq \Delta\}$ punktu k^* wskazanego przez krzyżową walidację, można policzyć następująco*

$$P\Big(\arg\min_{k\in\{1,\ldots,K\}} er_P(\widehat{f_{k,m}}) \in \{k\colon |k-k^*|\leq\Delta\}\Big) = \sum_{k\in\{k\colon |k-k^*|\leq\Delta\}}\int_{-\infty}^{\infty}\Big(\prod_{l\neq k}\int_{r_k}^{\infty} p_l(r_l)dr_l\Big)p_k(r_k)dr_k, \tag{43}$$

gdzie p_l, p_k są gęstościami normalnymi zdefiniowanymi jak w (41), o jednostajnym błędzie przybliżenia rzędu $O\Big(\Big(1+\frac{1}{\sqrt{n-1}}+\sqrt{n}\Big)\frac{1}{\sqrt{m}}\Big)$.

Dowody powyższych twierdzeń wykorzystują: Centralne Twierdzenie Graniczne, złożenia nierówności Chernoffa / Hoeffdinga oraz twierdzenie Berry'ego–Esséen'a dotyczące dokładności przybliżeń rozkładu sumy zmiennych losowych za pomocą rozkładów normalnych (Berry, 1941; Esséen, 1942; Shevtsova, 2007).

Dobrą ilustracją dla twierdzeń jest Rys. 2b. Na rysunku konkretna realizacja wyników krzyżowej walidacji, która miała miejsce w eksperymencie, wskazuje na punkt $k^* = 5$ jako punkt optymalnej złożoności. Nad poszczególnymi punktami struktury narysowano gęstości rozkładów prawdopodobieństwa na nieznany błąd prawdziwy $\mathrm{er}_P(\widehat{f_{k,m}})$, o których mowa w twierdzeniach. Na podstawie tych właśnie rozkładów twierdzenia pozwalają wyznaczyć interesujące nas prawdopodobieństwo rozbieżności punktów minimum dla danych warunków eksperymentu.

[15] W sensie supremum po wszystkich r dla funkcji dystrybuanty.

Wyjaśnienia wymaga termin *minimalne prawdopodobieństwo* pojawiający się w twierdzeniach. Składają się na niego dwa powody. Po pierwsze, obliczenia prawdopodobieństw oparte są na rozkładach będących przybliżeniami nieznanych rozkładów, z których tak naprawdę wolelibyśmy korzystać. Niemniej jednak przybliżenia te dobrane są w ten sposób (poprzez nadmiarowe wariancje), że co najwyżej pogarszamy wynikową miarę prawdopodobieństwa. Formalnie: dla ustalonego $0 < \delta < 1$ i dwóch bliskich rozkładów A_*, A o gęstościach p_{A_*}, p_A, mówimy, że A_* jest *pesymistycznie* przybliżany przez A, wtedy i tylko wtedy, gdy dla wszystkich kwantyli $a_{1-\delta_0/2}$, gdzie $\delta_0 \leqslant \delta$, wziętych z A zachodzi warunek

$$\int_{-a_{1-\delta_0/2}}^{a_{1-\delta_0/2}} p_{A_*}(x)dx \geqslant \int_{-a_{1-\delta_0/2}}^{a_{1-\delta_0/2}} p_A(x)dx. \tag{44}$$

Po drugie, udowodniono pomocniczo (Klęsk, 2010*b*, twierdzenie 4), że poprzez zawężenie nadmiarowej wariancji dowolnego przybliżającego rozkładu (dla dowolnej pozycji struktury) można co najwyżej podnieść wynikową miarę prawdopodobieństwa zdarzenia, że k^* wskazuje prawdziwe minimum, a nie pomniejszyć.

Patrząc jakościowo na wyznaczone gęstości (41) i ich stałe (42), można zauważyć, że ze względu na rozmiar próby m wariancja maleje się wraz z czynnikiem $1/\sqrt{m}$. Jest to intuicyjny rezultat probabilistyczny. Ze względu na liczbę n iteracji walidacyjnych, wariancja skaluje się według czynnika $1 + 1/\sqrt{n-1} + \sqrt{n}$. Widać, że w szczególności walidacja *leave-one-out* silnie powiększa wariancję. Oczywiście, aby uzyskać wysoki wynik interesującego nas prawdopodobieństwa (które przychylałoby się na rzecz wskazania k^* jako optymalnego), należałoby dobrać właściwą kombinację liczb m i n, tak aby wystarczająco zawęzić wariancję. Należy zaznaczyć, że oprócz opisanego wpływu powyższych stałych na wynikowe prawdopodobieństwo, mają one bardzo podobny wpływ na dokładność przybliżeń normalnych (CTG), patrz (Klęsk, 2010*b*, dodatek B).

Oprócz opisanego tu scenariusza — porównanie minimum wskazanego przez krzyżową walidację z minimum błędu prawdziwego — w pracy (Klęsk, 2010*b*) rozpatrzone są jeszcze dwa inne: (1) porównanie minimum wskazanego przez ograniczenia Vapnika z nieznanym minimum krzyżowej walidacji (gdy nie chcemy jej wyliczać z uwagi na koszt czasowy); (2) porównanie minimum ograniczeń Vapnika z nieznanym minimum błędów prawdziwych.

Od strony matematycznej, obliczenie rozpatrywanych prawdopodobieństw poprzez odpowiednią technikę całkowania może być postrzegane jako rozwiązanie bardziej elementarnego problemu, nie związanego z uczeniem maszynowym. Problem ten można postawić jako zadanie znalezienia optimum funkcji probabilistycznej (nie deterministycznej) określonej nad zbiorem skończonym.

Porównanie ciasności ograniczeń na błąd prawdziwy prowadzonych od dwóch wersji nierówności Chernoffa

Jak już wspomniano, jednym z narzędzi służących do wyprowadzenia ograniczeń na błąd prawdziwy jest nierówność Chernoffa. Oprócz tzw. postaci addytywnej (10) tej nierówności, istnieje także jej postać multiplikatywna (tu podana w wersji jednostronnej):

$$P_m\left(\frac{p - \nu_m}{\sqrt{p}} > \epsilon\right) \leqslant \exp(-\epsilon^2 m/2). \tag{45}$$

Vapnik (1998) podaje ciasne ilościowo ograniczenia wyprowadzane za pomocą obu tych wersji. Oto dwa wybrane przykłady dla nieskończonego zbioru funkcji F, wyrażone w terminach funkcji wzrostu G^F jako wybranego pojęcia pojemności zbioru funkcji.

Na podstawie addytywnej nierówności Chernoffa:

$$\mathrm{er}_P(\widehat{f}) \leqslant \widehat{\mathrm{er}}_{\mathbf{z}}(\widehat{f}) + \sqrt{\frac{\ln G^F(2m) - \ln(\delta/4)}{m}} + \frac{1}{m}. \tag{46}$$

Na podstawie multiplikatywnej nierówności Chernoffa:

$$\mathrm{er}_P(\widehat{f}) \leqslant \widehat{\mathrm{er}}_{\mathbf{z}}(\widehat{f}) + 2\frac{\ln G^F(2m) - \ln(\delta/4)}{m}\left(1 + \sqrt{1 + \frac{\widehat{\mathrm{er}}_{\mathbf{z}}(\widehat{f})m}{\ln G^F(2m) - \ln(\delta/4)}}\right). \tag{47}$$

Przypomnijmy, że wartość funkcji wzrostu można w powyższych ograniczeniach zastąpić wyrażeniem związanym z wymiarem VC za pomocą lematu Sauera.

W pracy (Klęsk, 2010*a*) dokonano analitycznego porównania powyższych ograniczeń, chcąc stwierdzić, kiedy w praktyce korzystniejsze jest wybranie jednego z nich. Podano między innymi następujące wyniki.

Twierdzenie 12 *(Klęsk, 2010*a*, twierdzenie 2) Niech F będzie nieskończonym zbiorem funkcji zerojedynkowych. Wtedy, korzystniej jest stosować ograniczenie (47) na błąd prawdziwy wyprowadzone na podstawie multiplikatywnej nierówności Chernoffa, raczej niż ograniczenie (46), jeżeli błąd na próbie uczącej dla wybranej funkcji $\widehat{f}$ jest mniejszy niż* $1/4$ *i następujące dwa warunki są spełnione:*

1. $m > 16(\ln G^F(2m) - \ln(\delta/4))/(1 - 4\widehat{er}_{\mathbf{z}}(\widehat{f}))^2$,
2. $G^F(2m) < \exp\left(m(1 - 4\widehat{er}_{\mathbf{z}}(\widehat{f}))^2/16 + \ln(\delta/4)\right)$.

Konsekwencja 2 Dla funkcji wzrostu G^F i rozmiaru próby m, ograniczenie (47) jest korzystniejsze, gdy błąd uczący $\widehat{\mathrm{er}}_{\mathbf{z}}(\widehat{f})$ jest mniejszy niż $1/4\left(1 - \sqrt{16(\ln G^F(2m) - \ln(\delta/4))/m}\right)$.

Wymiar Vapnika-Chervonenkisa dla algorytmu najbliższych sąsiadów

We wprowadzeniu na str. 9 podano wybrane przykłady wymiaru VC dla pewnych znanych zbiorów funkcji. Wspomniano również, że istnieją zbiory funkcji, dla których wymiar VC pozostaje nieznany.

Jednym z algorytmów, który nastręcza pewnych trudności w tej materii, jest dobrze znany algorytm ***k*-najbliższych sąsiadów** (ang. *k-Nearest Neighbors*). Po pierwsze, trudności wynikają z faktu, że ciężko jest o natychmiastowe wskazanie zbioru funkcji, z którego wybierany jest model. A należy zaznaczyć, że ściśle rzecz biorąc, wymiar VC jest własnością zbioru funkcji, nie zaś algorytmu uczącego. Po drugie, w literaturze spotkać można pewne stwierdzenia dające mylne przekonanie na temat tego algorytmu, w szczególności na przykład następujące zdanie z pracy (Kearns i Ron, 1999, punkt 3):

> "(...) dla algorytmów najbliższych sąsiadów nie istnieje ustalona 'klasa hipotez' o skończonym wymiarze VC — algorytm może wybierać dowolnie złożone hipotezy (...)".

Przyjęcie tego zdania za prawdziwe, oznaczałoby, że algorytm k-NN nie jest w stanie dobrze uogólniać. Uściślając, zdanie to jest prawdziwe w takim sensie, że algorytm 1-NN faktycznie implikuje nieskończony wymiar VC[16], a także że dla ustalonego $k > 1$ rozróżnialnych funkcji nadal może być dowolnie dużo, jeżeli będziemy podnosić rozmiar próby. Zdanie to *nie* jest jednak prawdziwe w takim sensie, iż rzekomo niemożliwym jest skonstruowanie zbioru funkcji (związanego z algorytmem NN), który

[16] Jak wiadomo algorytm 1-NN realizuje podział Voronoia na przestrzeni wejściowej.

posiadałby pewne stopnie swobody oraz miałby własność skończonego wymiaru VC. W pracy (Klęsk i Korzeń, 2011) autorom udało się uporządkować istniejące niejasności i podać przykład konstrukcji takiego zbioru.

Pomysł autorów opiera się na prostym przeformułowaniu oryginalnego algorytmu, poprzez wprowadzenie ułamka $\alpha \in (0,1)$, reprezentującego **ustalony procent najbliższych sąsiadów**. Zmienianie wartości α będzie stanowić sterowanie złożonością modelu. Wychodząc od tego pomysłu, Klęsk i Korzeń (2011) najpierw definiują **algorytm α-NN***, następnie pokazują, jak skonstruować zbiór funkcji z nim związany, i wreszcie dowodzą, że **wymiar VC tego zbioru jest skończony**. Dodatkowo udowodniono, że w pewnym szczególnym (ale typowo stosowanym w praktyce) przypadku wymiar ten wynosi dokładnie: $\lfloor 2/\alpha \rfloor$ — co można uznać za rezultat na skalę światową[17]. A zatem skoro zagwarantowany jest ograniczony wymiar VC, to tym samym zagwarantowana jest również zdolność do uogólniania na pewnym poziomie, co stanowi polemikę z cytowanym zdaniem z (Kearns i Ron, 1999). Oto najważniejsze elementy pracy (Klęsk i Korzeń, 2011).

Definicja 4 *Mając daną próbę uczącą o rozmiarze m, niech algorytm o nazwie α-NN*, $\alpha \in (0,1)$, będzie równoważny tradycyjnemu algorytmowi $k(\alpha,m)$-NN, gdzie liczba sąsiadów branych pod uwagę jest wyznaczana następująco:*

$$k(\alpha,m) = \begin{cases} \lceil \alpha m \rceil, & \text{gdy } \lceil \alpha m \rceil \text{ jest nieparzyste;} \\ \lceil \alpha m \rceil - 1, & \text{w przeciwnym razie.} \end{cases} \tag{48}$$

Jake widać, w sposób zamierzony k utrzymywane jest jako nieparzyste. To gwarantuje, że dla binarnej klasyfikacji wynikowa klasa, może zawsze być wybrana jednoznacznie jako klasa większościowa w zbiorze sąsiadów[18].

Następnym elementem jest zdefiniowanie *zbioru funkcji*, z którego wybierane są konkretne modele NN. W przypadku innych klasyfikatorów — liniowych, wielomianów, sieci neuronowych, itp. łatwiej jest dostrzec, z jakim zbiorem funkcji mamy do czynienia i jakie są jego stopnie swobody. Osoby stosujące algorytm najbliższych sąsiadów, zwykle natychmiastowo widzą *pojedynczą funkcję*, która stanowi wynik algorytmu uczącego, ale nie widzą z jakiego *zbioru* ona pochodzi. W pracy (Klęsk i Korzeń, 2011) pokazano, jak można patrzeć na interesujący nas zbiór i zdefiniowano jego stopnie swobody za pomocą: (1) **punktów odniesienia**, które mogą być dowolnie przemieszczane (i niekoniecznie muszą pokrywać się z punktami danych) oraz (2) **etykiet** dla punktów odniesienia, które również mogą być zmieniane.

Wprowadźmy pomocniczą notację reprezentującą zbiór najbliższych sąsiadów. Zakładamy, że pewna metryka ρ została ustalona. Mając dany zbiór tzw. *punktów odniesienia*: $\{\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_M\}$ rozmieszczonych w przestrzeni wejściowej, $\mathbf{c}_j \in \mathbb{R}^d$, niech

$$k\text{-NN}(\mathbf{x}) = \{n_1, n_2, \ldots, n_k\} \tag{49}$$

oznacza zbiór k indeksów tych punktów odniesienia $\mathbf{c}_j$, które są najbliższe do $\mathbf{x}$ w metryce ρ. Oznacza to, że dla każdego indeksu $n \in k\text{-NN}(\mathbf{x})$ i każdego indeksu $\overline{n} \in \{1,\ldots,M\} \setminus k\text{-NN}(\mathbf{x})$ mamy $\rho(\mathbf{x},\mathbf{c}_n) \leq \rho(\mathbf{x},\mathbf{c}_{\overline{n}})$.

[17] Warto tu wspomnieć, że istniały wcześniej w literaturze oszacowania na *efektywną liczbę parametrów* algorytmu najbliższych sąsiadów, mówiące, że wynosi ona m/k (Hastie, Tibshirani i Friedman, 2009; Jahne, 2004; Devroye, Gyorfi i Lugosi, 1996). Wynik z (Klęsk i Korzeń, 2011) wskazuje, że liczba ta jest raczej proporcjonalna do $2m/k$, zważywszy na fakt, że dla ustalonego m, wartość α można interpretować jako stosunek k/m.

[18] Inny możliwy problem ma miejsce wtedy, gdy sam zbiór sąsiadów nie pozwala się wybrać jednoznacznie, t.j. gdy mamy więcej niż k punktów mających taką samą odległość do danego punktu — np. wierzchołki czworościanu foremnego gdy $k = 3$. Dla przestrzeni ciągłych prawdopodobieństwo tego typu konfliktu wynosi zero, ale dla przestrzeni dyskretnych jest dodatnie i wówczas należy wybrać sąsiedztwo w sposób arbitralny.

Definicja 5 *Dla ustalonego $\alpha \in (0,1)$ i liczby naturalnej M, niech $F_{\alpha,M}$ oznacza następujący zbiór funkcji zerojedynkowych (skojarzony z algorytmem najbliższych sąsiadów):*

$$F_{\alpha,M} = \left\{ f_{\alpha,M}(\mathbf{x}; \mathbf{c}_1, \ldots, \mathbf{c}_M, t_1, \ldots, t_M) \right\}, \tag{50}$$

gdzie $\mathbf{c}_j \in \mathbb{R}^d$ są punktami odniesienia a $t_j \in \{0,1\}$ są etykietami klas skojarzonymi z punktami odniesienia. Postać funkcyjna dla f to:

$$f_{\alpha,M}(\mathbf{x}; \mathbf{c}_1, \ldots, \mathbf{c}_M, t_1, \ldots, t_M) = \left\lfloor \sum_{j \in\, k(\alpha,M)\text{-}NN(\mathbf{x})} t_j \Big/ \left\lceil \frac{1}{2} k(\alpha, M) \right\rceil \right\rfloor. \tag{51}$$

Liczba parametrów tego zbioru to $M(d+1)$. Sposób wyliczania wartości funkcji (51) jest równoważny tradycyjnemu głosowaniu większościowemu z sąsiedztwa.

Na podany zbiór warto patrzeć w ten sposób, że zmieniając pozycje punktów odniesienia $\mathbf{c}_j$ i dodatkowo iterując po wszystkich 2^M możliwych ustawieniach dla etykiet t_j, generujemy wszystkie możliwe kształty granicy decyzyjnej. Należy też zwrócić uwagę, że skoro rozważamy teraz zbiór funkcji (a nie konkretny problem uczenia), definicja zbioru $F_{\alpha,M}$ jest niezależna od próby. Stąd też zapis $k(\alpha, M)$ zamiast $k(\alpha, m)$. Jednakże, dalszą intencją jest to, że gdy dana już będzie próba ucząca, wówczas stopnie swobody $\mathbf{c}_1, \ldots, \mathbf{c}_M, t_1, \ldots, t_M$ zostaną nastrojone tak, aby najlepiej dopasować się do niej.

Autorzy pokazali dwa możliwe sposoby uczenia dla zdefiniowanego zbioru $F_{\alpha,M}$. Pierwszy jest równoważny powszechnemu rozumieniu klasyfikatora NN i polega nastawieniu stopni swobody wprost na dane — tzn. ustalając $M := m$ sięgamy po zbiór $F_{\alpha,m}$ i ustawiamy $\mathbf{c}_j := \mathbf{x}_j,\ t_j := \mathbf{y}_j, \quad j = 1, \ldots, m$. Drugi zaproponowany sposób pozwala wybrać $M < m$ i wykorzystuje klasteryzację (Klęsk i Korzeń, 2011, punkt 4.2). Można zadać pytanie: czy pierwszy sposób (tradycyjny) zgodnie z powszechnym przekonaniem gwarantuje minimalizację błędu na próbie dla $k(\alpha, m) > 1$? Co ciekawe odpowiedź jest przecząca, i nietrudno o wskazanie kontrprzykładu (Klęsk i Korzeń, 2011, punkt 4.1).

Następujące twierdzenie stanowi główny rezultat cytowanej pracy.

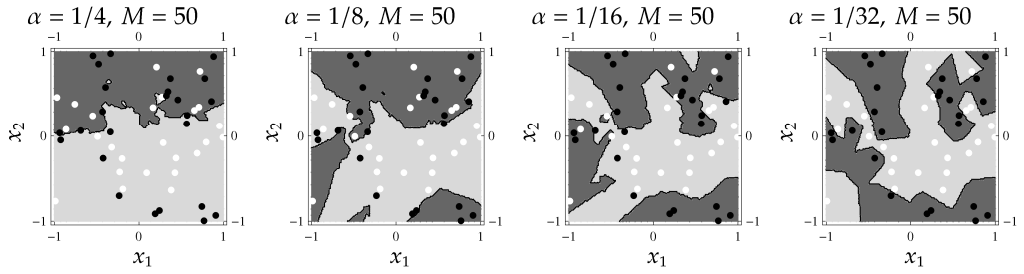
Twierdzenie 13 *(Klęsk i Korzeń, 2011, twierdzenie 1) Wymiar Vapnika-Chervonenkisa dla zbioru $F_{\alpha,M}$ jest skończony i prawdziwe są następujące zdania:*

$$VC\text{-}dim(F_{\alpha,M}) = M, \qquad dla\ M \leqslant \lfloor 2/\alpha \rfloor; \tag{52}$$

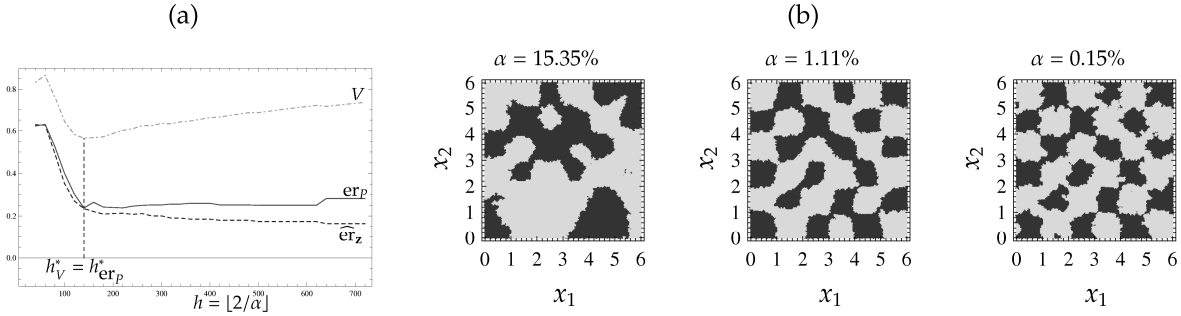
$$\lfloor 2/\alpha \rfloor \leqslant VC\text{-}dim(F_{\alpha,M}) < M, \qquad dla\ M > \lfloor 2/\alpha \rfloor. \tag{53}$$

Konsekwencja 3 Nie istnieje próba $\mathbf{z}_1, \ldots, \mathbf{z}_m$ o rozmiarze $m = \lfloor 2/\alpha \rfloor + 1$, która byłaby roztrzaskiwana przez zbiór funkcji $F_{\alpha,\lfloor 2/\alpha \rfloor + 1}$.

Oprócz wyników teoretycznych, pokazano także szereg praktycznych eksperymentów realizowanych w stylu procedury SRM. Sterując ułamkiem α wybierano optymalną złożoność — poglądowe ilustracje przedstawiono na rysunkach 3 i 4. W szczególności możliwe jest utworzenie ciągu zbiorów funkcji $F_{\alpha_1,M}, F_{\alpha_2,M}, F_{\alpha_2,M}, \ldots$, dla których wartość wymiaru VC przyrastałaby dokładnie o jeden, jeżeli ciąg ułamków dobierzemy następująco $(\alpha_1, \alpha_2, \alpha_3, \ldots) = (2/3, 2/4, 2/5, \ldots)$.



Rysunek 3: Ilustracja wzrastającej pojemności zbiorów $F_{\alpha,M}$ wraz z obniżaniem ułamka α. Dla lepszego wyobrażenia na rysunku punkty odniesienia i ich etykiety celowo pozostają ustalone.



Rysunek 4: Klasyfikacja dla wzorca „szachownica". (a) Przykładowy wynik procedury SRM. (b) Przykładowe trzy wynikowe modele: niewystarczająco złożony, odpowiednio dobrze złożony, zbyt złożony (dopasowuje się do szumów).

Probabilistyczne szacowanie wymiaru Vapnika-Chervonenkisa

Ustalenie wymiaru VC pewnych zbiorów funkcji wymaga formalnych dowodów. Są to zwykle trudne dowody zawierające elementy geometryczne i kombinatoryczne. Przykładem badań habilitanta w tym obszarze była praca (Klęsk i Korzeń, 2011) przedstawiona w poprzednim punkcie. Trudności napotykane przy próbach konstrukcji takich dowodów są motywacją do szukania podejść alternatywnych. I tak, naturalnym następstwem wspomnianych badań są badania habilitanta nad „miękkim" podejściem do tego zagadnienia, polegającym na szacowaniu wymiaru VC (bez dowodu).

W pracy (Klęsk, 2012)[19] przedstawiono pomysł na **algorytm, który dla podanego na wejście dowolnego zbioru funkcji (i algorytmu uczącego) szacuje wymiar VC z pewną zadaną precyzją probabilistyczną**. Zadana precyzja mówi, jak często i o ile znalezione oszacowanie może różnić się od dokładnej nieznanej wartości wymiaru VC. W pracy przedstawiono dwa warianty zaproponowanego algorytmu — nazwane jako A i A'. Przedstawiono analizę zbieżności tych algorytmów i złożoność obliczeniową. Wyjaśniając skrótowo, oba algorytmy pracują wg podejścia, które można by opisać jako *mnóż lub dziel i zwyciężaj*. Oznacza to, że dla różnych rozmiarów prób sprawdzana jest możliwość roztrzaskiwania. Rozmiary próby są podwajane (*mnóż*) lub połowione (*dziel*), co przekłada się na **logarytmiczny czas przeszukiwań ze względu na zadaną precyzję**.

Na potrzeby zaproponowanych algorytmów habilitant zdefiniował kilka pojęć, które mogą być

[19]Praca jest przyjęta na konferencję ICAART w Portugalii, która odbędzie się w terminie: 6–8 lutego, 2012 r. Habilitant będzie prezentował pracę w ramach 20 minutowego seminarium. Praca otrzymała trzy pozytywne recenzje w tzw. systemie 'blind'. Decyzję o przyjęciu załączono do niniejszego wniosku. Nie są natomiast jeszcze znane takie szczegóły publikacyjne jak: numery stron, numer tomu materiałów konferencyjnych, w których ukaże się praca.

rozumiane jako probabilistyczne odpowiedniki znanych pojęć twardych i dają ogólną intuicję.

Definicja 6 *Mówimy, że* $\mu^F(m)$ *jest* ***miarą roztrzaskiwalności*** *obliczoną względem rozkładu* P^m *jako*

$$\mu^F(m) = \int_{\mathbf{Z}^m} [\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m] dP^m(\mathbf{z}_1,\ldots,\mathbf{z}_m). \tag{54}$$

Intuicyjnie, *miara roztrzaskiwalności* mówi, jak często pojawiają się próby czerpane z P^m, które można roztrzaskać za pomocą funkcji z F. Pojęcie to dobrze jest rozważać w parze z pojęciem funkcji wzrostu $G^F(m)$ (patrz str. 9). Wyobraźmy sobie pewną metodę, która próbuje wykryć argument $\mathbf{z}_1,\ldots,\mathbf{z}_m$ w P^m, dla którego osiągane jest supremum liczby rozróżnialnych funkcji. Dla ścisłości należy przypomnieć, że definicja $G^F(m)$ jest niezależna od rozkładu, a po drugie nawet jeśliby była ona zależna od rozkładu, to supremum może być osiągane na zbiorze o mierze zero. Niemniej jednak prawdziwa jest intuicja, że im mniejsze $\mu^F(m)$ tym trudniej jest wskazać supremum reprezentowane przez $G^F(m)$. W szczególności jeżeli $G^F(m) < 2^m$, to na pewno $\mu^F(m) = 0$.

Definicja 7 *Mówimy, że zbiór* F *(funkcji zerojedynkowych) jest* m*-roztrzaskujący ze względu na rozkład* P^m *(lub: roztrzaskuje pewne próby o rozmiarze* m *z rozkładu* P^m*) jeżeli* $\mu^F(m) > 0$.

Definicja 8 *Mówimy, że zbiór* F *nie jest* m*-roztrzaskujący ze względu na rozkład* P^m *wszędzie, jeżeli zachodzą dwa warunki: (1)* $\mu^F(m) = 0$*, (2)* $\nexists\, \mathbf{z}_1,\ldots,\mathbf{z}_m$ *t.ż.* $\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m$.

Definicja 9 *Mówimy, że zbiór* F *nie jest* m*-roztrzaskujący ze względu na rozkład* P^m *prawie wszędzie, jeżeli zachodzą dwa warunki: (1)* $\exists\, \mathbf{z}_1,\ldots,\mathbf{z}_m$ *t.ż.* $\#(l_F)_{|\mathbf{z}_1,\ldots,\mathbf{z}_m} = 2^m$*, (2)* $\mu^F(m) = 0$.

Poniżej przedstawiono podstawowy wariant algorytmu o nazwie A. W ramach wewnętrznej pętli algorytm ten wywołuje pomocniczy algorytm B. Wynikiem działania algorytmu A jest liczba naturalna,

Algorytm $A_{\epsilon,\delta}(F, P)$

1. Ustaw $m_L := 1$, $m_U := \infty$, $m := m_L$.
2. Powtarzaj dopóki $m_U - m_L > 1$:
   - 2.1. Ustaw $s := 0$.
   - 2.2. Powtarzaj $n = \lceil -\ln\delta/(2\epsilon^2)\rceil$ razy:
     - 2.2.1 Zaczerpnij próbę $\mathbf{z}_1,\ldots,\mathbf{z}_m$ z rozkładu P^m.
     - 2.2.2 Jeżeli $B(F;\mathbf{z}_1,\ldots,\mathbf{z}_m) = 1$ wówczas ustaw $s := 1$ i wyskocz z pętli 2.2.
   - 2.3 Jeżeli $m_U = \infty$:
     - 2.3.1 Jeżeli $s = 1$, to ustaw $m_L := 2m$, $m := m_L$.
     - 2.3.2 W przeciwnym razie ustaw $m_L := 1/2m$, $m_U := m$, $m := (m_L + m_U)/2$.
   - 2.4 W przeciwnym razie:
     - 2.4.1 Jeżeli $s = 1$ to ustaw $m_L := m$, $m := (m_L + m_U)/2$.
     - 2.4.2 W przeciwnym razie $m_U := m$, $m := (m_L + m_U)/2$.
3. Zwróć $\lfloor m_L \rfloor$.

Algorytm pomocniczy $B(F; \mathbf{z}_1, \ldots, \mathbf{z}_m)$ **do sprawdzenia roztrzaskiwania**

1. Dla wszystkich $(t_1, \ldots, t_m) \in \{0,1\}^m$:
   1.1. Utwórz tymczasową próbę uczącą $\mathbf{z} = (x_1, t_1), \ldots, (x_m, t_m)$ i uruchom algorytm uczący L na niej, otrzymując jako wynik $\widehat{f}$.
   1.2. Jeżeli $\widehat{\mathrm{er}}_{\mathbf{z}}(\widehat{f}) > 0$ zwróć 0.
2. Zwróć 1.

którą można nazwać **probabilistycznym wymiarem Vapnika-Chervonenkisa**. Stanowi ona dolne oszacowanie na prawdziwy wymiar VC. Jeżeli liczba ta wynosi h, to wiemy, że na pewno istnieje próba o rozmiarze h, która jest roztrzaskiwana przez F, a także wiemy, że z prawdopodobieństwem $1 - \delta$ lub większym $\mu^F(h+1) < \epsilon$. Innymi słowy, z dużym prawdopodobieństwem nie istnieją próby o rozmiarze $h+1$, które byłyby roztrzaskiwalne.

Okazuje się, że zbieżność algorytmu A wygodnie jest rozważać w terminach miar roztrzaskiwalności dla danego problemu. Rozważmy ciąg $\mu_1 = \mu^F(1), \mu_2 = \mu^F(2), \ldots$ zbudowany dla wzrastającego rozmiaru próby. Po pierwsze warto zaobserwować, że jest to ciąg nierosnący (Klęsk, 2012, lemat 1). Po drugie interesującą obserwacją jest to, że w ogólności można wyróżnić dwie duże rodziny zbiorów funkcji F — rodzinę, dla której wspomniany ciąg składa się *tylko* z jedynek i zer tzn. ciąg ma wzorzec $(1, 1, \ldots, 1, 0, \ldots)$; oraz rodzinę, dla której ciąg ten zawiera ułamki. Dłuższe zastanowienie pozwala też zauważyć, że typowo stosowane w uczeniu zbiory funkcji, gdzie bazami są np. swobodne płaszczyzny, kule, czy kostki należą do pierwszej rodziny. Poniższe twierdzenie podaje dokładny rozkład prawdopodobieństwa na zbieżność algorytmu A.

Twierdzenie 14 *Przypuśćmy, że $\mu_1, \mu_2, \ldots$ jest ciągiem miar roztrzaskiwalności dla danego zbioru funkcji F i rozkładu P. Niech $q = \lfloor \log_2 h \rfloor$ i niech $(h_q, h_{q-1}, \ldots, h_0)_2$ oznacza dwójkowy zapis każdej liczby naturalnej $h > 0$. Wtedy, rozkład prawdopodobieństwa na wyniki, do których algorytm A może zbiegać jest następujący*

$$\begin{aligned} p(0) &= (1-\mu_1)^n, \\ p(1) &= \Big(1-(1-\mu_1)^n\Big)(1-\mu_2)^n, \\ p(h) &= \prod_{k=0}^{q}\Big(1-(1-\mu_{2^k})^n\Big)(1-\mu_{2^{q+1}})^n \cdot \prod_{k=0}^{q-1}\Big(h_{q-k-1} + (-1)^{h_{q-k-1}}(1-\mu_{i(h,k)})^n\Big), \end{aligned} \tag{55}$$

dla $h \geq 2$, gdzie

$$i(h,k) = \frac{1}{2}(2^{q+1} + 2^q) + \sum_{j=1}^{k}(-1)^{1-h_{q-j}} \cdot 2^{q-j-1}. \tag{56}$$

Konsekwencja 4 Przypuśćmy, że VC-dim$(F) = h^*$ oraz że ciąg miar roztrzaskiwalności $\mu_1, \mu_2, \ldots$ składa się tylko z jedynek i zer. Wówczas algorytm A zbiega z prawdopodobieństwem jeden do poprawnego wyniku, t.j. $p(h^*) = 1$ i $p(h) = 0$ dla wszystkich $h \neq h^*$ i dla wszystkich wyborów $0 < \epsilon, \delta < 1$.

W pracy (Klęsk, 2012) opisano także wzmiankowany tu wariant algorytmu o nazwie A', który stanowi usprawnienie algorytmu A w sensie kosztowności obliczeniowej. Wariant A' wprowadza oszczędniejszy algorytm pomocniczy do sprawdzenia roztrzaskiwania (proszę zwrócić uwagę, że oryginalny algorytm B wymagał wykonania 2^m iteracji). Dzieje się to poprzez wprowadzenie podwójnej precyzji probabilistycznej w formie $(\epsilon_1, \delta_1, \epsilon_2, \delta_2)$. W efekcie złożoność algorytmu A' jest w pełni logarytmiczna.

Wynikiem algorytmu A' liczba, która może być zarówno przeszacowaniem jak i niedoszacowaniem wymiaru VC. Oczywiście z zadanym (odpowiednio dużym) prawdopodobieństwem wynik ten będzie równy szukanemu wymiarowi VC.

Szacowanie maksymalnego marginesu separacji

Dla zbiorów danych liniowo-separowalnych znanym faktem jest to, że znalezienie płaszczyzny decyzyjnej o dużym **marginesie separacji** korzystnie wpływa na zdolność do uogólniania. Ta obserwacja była m.in. motywacją Vapnika do opracowania algorytmu SVM, który jest zorientowany na maksymalizowanie marginesu separacji. Warto także dodać, że za pomocą technicznej sztuczki z tzw. przekształceniem jądrowym (ang. *kernel transformation trick*), która przenosi problem do przestrzeni o wyższej wymiarowości, można łatwo poszukiwanie maksymalnego marginesu rozszerzyć na przypadek nieseparowalny liniowo (przykład na Rys. 6)

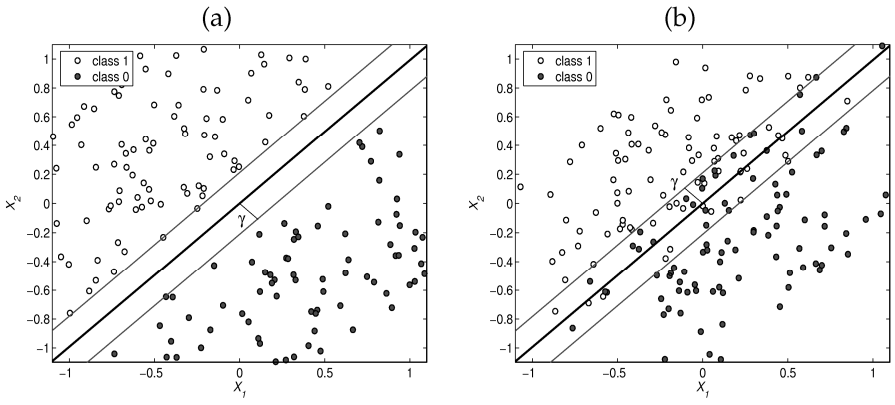
W pracy (Korzeń i Klęsk, 2008) autorzy rozważają klasyczny algorytm perceptronu Rosenblatt'a i jego związki z algorytmem SVM. Autorzy proponują dwie modyfikacje algorytmu Rosenblatt'a, które pozwalają na szacowanie maksymalnego marginesu separacji w danych. **Propozycja autorów po pierwsze poprawia zdolność do uogólniania klasycznego algorytmu Rosenblatt'a, a po drugie może być zastosowana pomocniczo do uczenia algorytmu SVM.** Oto ważniejsze elementy tej pracy.

Definicja 10 *Mówimy, że zbiór danych jest γ-separowalny (lub separowalny z marginesem γ) jeżeli:*

$$\exists \mathbf{w} \quad \forall i \in \{1,\ldots,m\} \qquad \frac{y_i}{\|\mathbf{w}'\|}, \mathbf{w}^T \mathbf{x}_i \geq \gamma, \tag{57}$$

gdzie $\mathbf{w} = (w_0, w_1, \ldots, w_d)$, $\mathbf{w}' = (w_1, \ldots, w_d)$ $\mathbf{x}_i = (1, x_1, \ldots, x_d)$ *oraz* $y_i \in \{-1, 1\}$ *(etykiety klas).*

Warto zauważyć, że w powyższej definicji **margines γ może być w szczególności ujemny**, patrz Rys. 5.



Rysunek 5: Ilustracja prostych z dodatnim (a) i ujemnym (b) marginesem separacji.

Liczbową wartość optymalnego (największego) marginesu można zdefiniować jako:

$$\gamma^* = \sup_{\|\mathbf{w}\|=1} \min_{i \in \{1,\ldots,m\}} \frac{y_i}{\|\mathbf{w}'\|} \mathbf{w}^T \mathbf{x}_i. \tag{58}$$

Na mocy twierdzenia Weierstrassa można łatwo pokazać, że γ^* zawsze istnieje — każdy zbiór danych jest separowalny z pewnym marginesem.

Niech $\mathbf{w}(k)$ oznacza zawartość wektora wag w k-tym kroku algorytmu. W klasycznym algorytmie Rosenblatt'a aktualizacja wag odbywa się według wzoru $\mathbf{w}(k+1) := \mathbf{w}(k) + \eta y_i \mathbf{x}_i$ (gdzie $\eta \in (0, 1]$ to współczynnik szybkości uczenia) i jest wykonywana pod warunkiem błędnego sklasyfikowania pary $(\mathbf{x}_i, y_i)$ dotychczasowym wektorem $\mathbf{w}(k)$. Warunek ten jest równoważny nierówności $y_i \mathbf{w}(k)^T \mathbf{x}_i \leqslant 0$. W pracy (Korzeń i Klęsk, 2008) autorzy zastępują ten warunek nierównością:

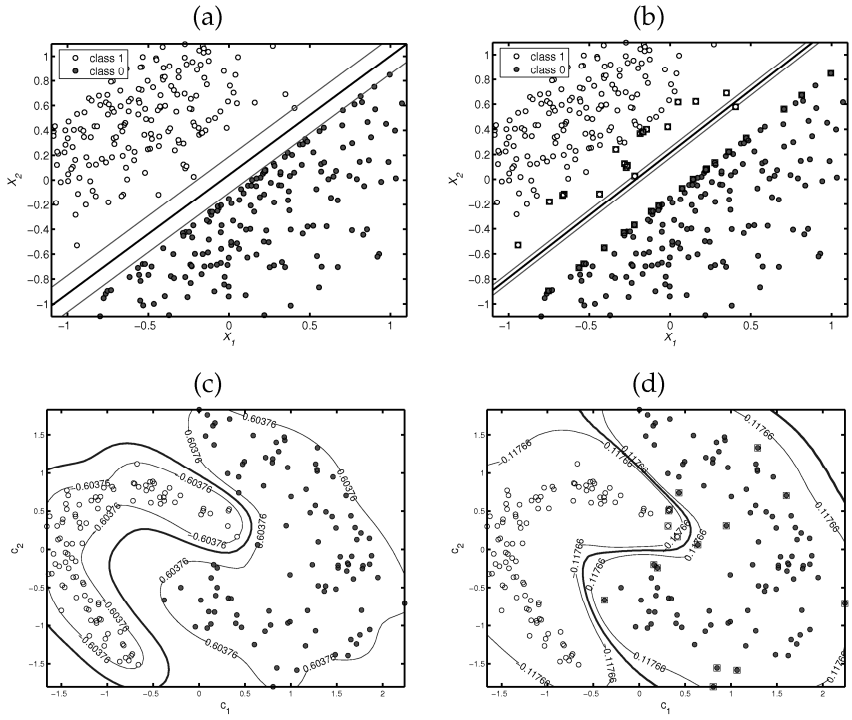
$$y_i \mathbf{w}(k)^T \mathbf{x}_i \leqslant \gamma(k) \|\mathbf{w}(k)'\|, \tag{59}$$

gdzie $\gamma(k)$ jest aktualnym marginesem. W przypadku, gdy $\gamma(k)$ jest dodatnie, oznacza to, że punkt danych musi być klasyfikowany nie tylko poprawnie, ale i oddalony o przynajmniej $\gamma(k)$ od płaszczyzny decyzyjnej. W przypadku, gdy $\gamma(k)$ jest ujemne, oznacza to, że punkt danych może być po złej stronie płaszczyzny decyzyjnej, ale w niegorszej odległości niż $|\gamma(k)|$. Idea algorytmu jest więc taka, że rozpoczynamy od odpowiednio dużego[20] $\gamma(0)$, które stopniowo obniżamy, aż do momentu, gdy wszystkie punkty danych spełnią nierówność (59). Wówczas ostatnie $\gamma(k)$ będzie stanowiło oszacowanie na optymalne γ^*. Autorzy proponują dwa sposoby obniżania wartości $\gamma(k)$:

$$\gamma(k+1) := \frac{y_i}{\|\mathbf{w}(k)'\|} \mathbf{w}(k)^T \mathbf{x}_i, \tag{60}$$

$$\gamma(k+1) := \gamma(k) + \alpha \left(\frac{y_i}{\|\mathbf{w}(k)'\|} \mathbf{w}(k)^T \mathbf{x}_i - \gamma(k) \right), \tag{61}$$

przekładające się odpowiednio na algorytmy o nazwach *varying margin perceptron* oraz *soft-varying margin perceptron*. Sposób pierwszy powoduje skokowe przestawienie marginesu na wartość, która od



Rysunek 6: Porównanie rozwiązań algorytmu Roseneblatt'a z marginesem γ (a, c) z rozwiązaniami SVM (b, d). Kwadratami zaznaczono punkty danych, które stanowią punkty podparcia w rozwiązaniu SVM.

razu pozwoli spełnić nierówność (59) dla aktualnego punktu danych $(\mathbf{x}_i, y_i)$. Sposób drugi stanowi

[20] $\gamma(0)$ można chociażby ustawić na promień danych.

miękkie przejście w kierunku tej wartości, z ułamkowym krokiem $\alpha \in (0, 1)$. Autorom udało się podać dowód zbieżności zmodyfikowanego algorytmu, ale tylko dla przypadku, gdy $\gamma(k)$ pozostaje dodatnie.

W pracy wykonano szereg praktycznych eksperymentów pokazujących, że zaproponowane algorytmy dobrze wykrywają maksymalny margines tkwiący w danych. Część eksperymentów porównywała otrzymane wyniki z wynikami algorytmu SVM — przykłady przedstawiono na Rys. 6.

Warto w tym miejscu przypomnieć postać kryterium optymalizacyjnego SVM. Należy zminimalizować wyrażenie $\|\mathbf{w}'\|^2 + C\sum_{i=1}^{m} \xi_i$ przy ograniczeniach $y_i(\mathbf{w}^T\mathbf{x}_i + \xi_i) \geqslant 1$ oraz $\xi_i \geqslant 0$ dla wszystkich $i = 1, \ldots, m$. Wielkości ξ_i reprezentują błędy (wpadanie punktów w pas marginesowy), a stała $C > 0$ steruje kompromisem pomiędzy marginesem a liczbą błędów. Jeżeli chcemy uzyskać duży margines kosztem większej liczby błędów, to należy dobrać małe C. Jest to dość wymagająca obliczeniowo optymalizacja kwadratowa z ograniczeniami szczególnie dla dużych zbiorów danych — liczba ograniczeń jest równa m. Dodatkowych trudności nastręcza stała C, dla której dobrą wartość wybiera się np. poprzez krzyżową walidację, co tym bardziej podnosi koszt obliczeniowy. Autorzy pracy (Korzeń i Klęsk, 2008) podają wskazówki, jak za pomocą oszacowanego maksymalnego marginesu dobrać wartość C bez krzyżowej walidacji. Warto zaznaczyć, że zaproponowane modyfikacje algorytmu Rosenblatt'a pozostają szybkimi algorytmami on-line o złożoności $O(m)$, takiej jaką ma algorytm klasyczny.

Literatura

Anthony, M. i Bartlett, P. (2009), *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, UK.

Bartlett, P. (1998), 'The sample complexity of pattern classification with neural networks: the size of weights is more important than the size of the network', *IEEE Transactions on Information Theory* **44**(2), 525–536.

Berry, A. (1941), 'The accuracy of the gaussian approximation to the sum of independent variates', *Trans. Amer. Math. Soc.* **49**, 122–136.

Chari, S., Rohatgi, P. i Srinivasan, A. (1994), Improved algorithms via approximations of probability distributions, *in* 'Proceedings of the Twenty-Sixth Annual ACM Symposium on the Theory of Computing', pp. 584–592.

Cherkassky, V. i Mulier, F. (1998), *Learning from data*, Adaptive and Learning Systems for Signal Processing, Communications and Control, John Wiley & Sons, inc.

Cybenko, G. (1989), Approximation by superpositions of sigmoids, *in* 'Mathematics of Control, Signals, and Systems', number 2, pp. 303–314.

Devroye, L., Gyorfi, L. i Lugosi, G. (1996), *A Probabilistic Theory of Pattern Recognition*, Springer Verlag, New York, Inc.

Esséen, C. (1942), 'On the Liapounoff limit of error in the theory of probability', *Ark. Mat. Astr. och Fys.* **28A**(9), 1–19.

Friedman, J. (1991), 'Multivariate adaptive regression splines', *Annals of Statistics* **19**(1), 1–67.

Hastie, T., Tibshirani, R. i Friedman, J. (2009), *The Elements of Statistical Learning*, Springer, New York, USA.

Haussler, D. (1995), 'Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension', *Journal of Combinatorial Theory, Series A* **69**(2), 217–232.

Haussler, D. i Long, P. (1995), 'A generalization of Sauer's lemma', *Journal of Combinatorial Theory, Series A* **71**(2), 219–240.

Jahne, B. (2004), *Practical Handbook on Image Processing for Scientific and Technical application*, CRC Press LLC.

Kearns, M. i Ron, D. (1999), 'Algorithmic stability and sanity-check bounds for leave-one-out cross-validation', *Neural Computation* **11**, 1427–1453.

Klęsk, P. (2010*a*), 'A comparison of certain generalization bounds of learning machines for practical applications', *Metody Informatyki Stosowanej* **2**(24), 35–45. Polska Akadamia Nauk oddział w Gdańsku, Szczecin, Poland.

Klęsk, P. (2010*b*), 'Probabilities of discrepancy between minima of cross-validation, Vapnik bounds and true risks', *International Journal of Applied Mathematics and Computer Science* **20**(3), 525–544. Zielona Góra, Poland.

Klęsk, P. (2011), A Relationship Between Cross-Validation and Vapnik Bounds on Generalization of Learning Machines, *in* 'Proceedings of the 3-rd International Conference on Agents and Artificial Intelligence — ICAART 2011', Vol. 1, SciTePress, Rome, Italy, pp. 5–17.

Klęsk, P. (2012), Probabilistic Estimation of Vapnik-Chervonenkis Dimension, *in* 'Proceedings of the 4-th International Conference on Agents and Artificial Intelligence — ICAART 2012', SciTePress, Vilamoura, Portugal.

Klęsk, P. i Korzeń, M. (2011), 'Sets of approximating functions with finite Vapnik-Chervonenkis dimension for nearest neighbors algorithms', *Pattern Recognition Letters* **32**(14), 1182–1893. Elsevier, New York, USA.

Kohavi, R. (1995), A study of cross-validation and boostrap for accuracy estimation and model selection, *in* 'International Joint Conference on Artificial Intelligence (IJCAI)'.

Korzeń, M. i Klęsk, P. (2008), Maximal Margin Estimation with Perceptron-like Algorithm, Vol. 5097/2008 of *Lecture Notes in Artificial Intelligence*, Springer, Berlin / Heidelberg, Germany, pp. 597–608. 9th International Conference on Artificial Intelligence and Soft-Computing — ICAISC 2008, Zakopane, Poland.

Quinlam, J. (1993), *Programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, USA.

Rumelhart, D., Hinton, G. i Williams, R. (1986), *Learning internal representations by error propagation*, Vol. 1 of *Parallel distributed processing: explorations in the microstructure of cognition*, MIT Press Cambridge, Massachusetts, USA.

Sauer, N. (1972), 'On the density of families of sets', *Journal of Combinatorial Theory, Series A* **13**, 145–147.

Shevtsova, I. (2007), 'Sharpening of the upper bound of the absolute constant in the Berry–Esséen inequality', *Theory of Probability and its Applications* **51**(3), 549–553.

Steel, J. (1978), 'Existence of submatrices with all possible columns', *Journal of Combinatorial Theory, Series A* **24**, 84–88.

Valiant, L. (1984), 'A theory of the learnable', *Communications of the ACM* **27**(11).

Vapnik, V. (1995), *The Nature of Statistical Learning Theory*, Springer Verlag, New York.

Vapnik, V. (1998), *Statistical Learning Theory: Inference from Small Samples*, Wiley, New York.

Vapnik, V. i Chervonenkis, A. (1971), 'On the uniform convergence of relative frequencies of events to their probabilities', *Theory of Probability and its Applications* **16**(2), 264–280.

Zhang, T. (2002), 'Covering number bounds of certain regularized linear function classes', *Journal of Machine Learning Research* **2**, 527–550.

5 Omówienie pozostałych osiągnięć naukowo–badawczych

5a Projekty badawcze

L.p.	Okres	Nazwa i dane projektu	Rola habilitanta	Budżet
1	2010–2012	Projekt badawczy własny MNiSW pt.: *"Algorytmy do oceny zdolności do uogólniania maszyn uczących się w terminach Statystycznej Teorii Uczenia Vapnika"*. Nr: N N516 424938.	**kierownik projektu i główny wykonawca**	136 tys. zł
2	2008–2011	Projekt międzynarodowy pt.: *"Measuring Interaction Between Quality of Life, Children Well-Being, Work and Public Policies"* prowadzony przez Uniwersytet w Modenie (Włochy) finansowany przez: Fondazione Cassa di Risparmio di Modena e Reggio Emilia, oraz Foundation Emmano Gorrieri for the Social Studies. (http://www.capp.unimo.it/ricerche/gender/childEN.html)	**członek zespołu, wykonawca**	80 tys. euro

Tablica 1: Projekty badawcze z udziałem habilitanta.

5b Nagrody i wyróżnienia

Przedstawiony spis nagród jest uporządkowany chronologicznie od nagród najnowszych do nagród najstarszych.

1. **Nagroda Rektora ZUT w Szczecinie** indywidualna III stopnia za osiągnięcia naukowe, 2011 r.

2. **Nagroda "Best Paper Award" w Rzymie** za najlepszy artykuł w kategorii „sztuczna inteligencja" na **międzynarodowej konferencji ICAART 2011** — 28–30 stycznia 2011 r., Rzym, Włochy. Kopia certyfikatu nagrody znajduje się w załączniku nr 6. Oficjalną informację o nagrodzie można również znaleźć na stronie: `http://www.icaart.org/previous_awards.asp`.

   Opis: *Tytuł nagrodzonego artykułu: "A Relationship Between Cross-Validation and Vapnik Bounds on Generalization of Learning Machines". Na konferencję nadesłano 367 artykułów. Każdy oceniany był przez trzech recenzentów w tzw. systemie 'blind' (artykuły przesyłane bez danych osobowych autorów). Proces recenzji przeszło pozytywnie 113 prac (31%). Artykuły były przyporządkowywane do trzech kategorii: poster, short paper, full paper. Do najlepszej kategorii full paper wybrano 32 prace, w tym artykuł habilitanta. W Rzymie habilitant przedstawiał pracę w ramach 30 minutowego seminarium. Oceny recenzentów oraz oceny słuchaczy-organizatorów złożyły się na ostateczny ranking prac, w wyniku którego artykułowi dra Klęska przyznano I miejsce w kategorii „sztuczna inteligencja" (jedna z dwóch kategorii).*

3. **Nagroda Dziekana Wydziału Informatyki ZUT w Szczecinie** za I miejsce w konkursie *Dean Cup*, 2010 r.

4. **Nagroda Prezydenta miasta Szczecina za najlepszą pracę magisterską dyplomanta** Bartosza Bielskiego p.t. „Pasywna identyfikacja systemów operacyjnych za pomocą sztucznych sieci neuronowych" realizowaną pod kierunkiem dra Klęska, maj 2009 r.

   Patrz także: `http://www.szczecin.eu/baltic_neopolis/aktualnosci/najlepsi_naukowcy.html_0`.

5. **II miejsce w konkursie na rozwiązanie problemu metalurgicznego** przeprowadzonego przy **międzynarodowej konferencji ICAISC 2008 w Zakopanem** — 22–26 czerwca 2008 r., Zakopane, Polska. Kopia nagrody znajduje się w załączniku nr 7.

   Opis: *Zadanie konkursowe polegało na opracowaniu (dowolnymi metodami) i zaprogramowaniu maszyny do regresji na podstawie danego zbioru 5 tys. przykładów dotyczących procesu wytopu stali. Każdy przykład opisany był poprzez 71 zmiennych wejściowych. W zbiorze występowały braki danych a także punkty odstające. Celem maszyny do regresji było przybliżenie z jak najmniejszym błędem 11 wielkości wyjściowych składających się na opis chemicznej zawartości powstałej stali.*

6. **Nagroda Rektora ZUT w Szczecinie**, indywidualna III stopnia za osiągnięcia naukowe, 2008 r.

7. **Nagroda Dziekana Wydziału Informatyki ZUT Szczecinie** za I miejsce w konkursie *Dean Cup*, 2008 r.

8. **Nagroda Rektora Politechniki Szczecińskiej**, indywidualna III stopnia w kategorii „młody pracownik", 2007 r.

9. **Nagroda Rektora Politechniki Szczecińskiej**, indywidualna II stopnia za osiągnięcia naukowe i uzyskanie stopnia doktora z wyróżnieniem, 2006 r.

5c Oryginalne osiągnięcia projektowe, konstrukcyjne lub technologiczne

System „*Contents Extractor*"

W 2010 r. na zlecenie firmy programistycznej Interactive Voice News sp. z o.o. (Al. Komisji Edukacji Narodowej 96/180, 02-722 Warszawa, `http://www.e-ivn.com`)[21] **habilitant zaprojektował i zaimplementował algorytm wykorzystujący uczenie maszynowe pozwalający na ekstrakcję właściwej treści artykułów na podstawie źródeł HTML stron internetowych pochodzących z portali informacyjnych**. System został pomyślnie **wdrożony** i pracuje po dziś dzień, patrz np. `http://www.i-talks.com`.

Firma Interactive Voice News chciała stworzyć oprogramowanie pozwalające na głosowe odczytywanie treści artykułów informacyjnych pochodzących z takich portali jak np. `gazeta.pl`, `wp.pl`, `onet.pl`, `tvn24.pl`, itp. W grę wchodziły wszelkie rodzaje artykułów informacyjnych — polityczne, geograficzne, sportowe, pogodowe, kulturalne, i inne. Jako docelowi odbiorcy planowanej aplikacji przewidziani zostali użytkownicy telefonów komórkowych (lub innych urządzeń mobilnych), którzy dzięki aplikacji mogliby posłuchać wiadomości np. w tramwaju lub podczas jazdy samochodem. Aplikacja była również planowania jako ułatwienie dla osób niewidomych.

Firma dysponowała gotowym modułem odpowiedzialnym za syntezę mowy na podstawie tekstu. Problemem, którego firma nie potrafiła rozwiązać, była ekstrakcja (rozpoznanie) właściwej treści artykułów ze źródeł HTML ze skutecznym wyeliminowaniem niechcianych elementów na stronie takich jak: menu, reklamy, komentarze, podpisy pod rysunkami, linki, forum dyskusyjne, itp. Elementy te są często celowo wplatane we właściwą treść artykułu przez twórców danego portalu.

Okazuje się, że powyższy problem nie daje się rozwiązać zadowalająco z użyciem technik czysto programistycznych, takich jak np. wyrażenia regularne czy analiza syntaktyczna. Po pierwsze próby tego typu rozwiązań dają niską skuteczność rozpoznawania właściwej treści. Wynika to z faktu, że bardzo trudno jest przewidzieć wszystkie możliwe sposoby prezentacji artykułu przez dany portal i rozwiązać zadanie pewnym „twardym" zestawem instrukcji *if/else* lub wzorców *regular expression*. Nie istnieją też gramatyki portali w sensie analizy syntaktycznej, które byłyby respektowane przez twórców portalu. Po drugie, jak wiadomo, portale internetowe z dużą częstotliwością zmieniają swoją szatę graficzną i układy stron, co praktycznie przekreśla wspomniane rozwiązanie, ponieważ wymagałoby ono gruntownego przeprogramowywania co pewien czas.

W związku z tym pojawiła się potrzeba rozwiązania opartego na sztucznej inteligencji, a dokładniej na uczeniu maszynowym. Zadanie postawiono jako problem klasyfikacji. Wyróżniono trzy klasy: *tytuł*, *treść*, *nie-treść*. Według projektu systemu, użytkownik-nauczyciel miałby uczyć klasyfikator przyrostowo poprzez wczytywanie adresów URL i znakowanie fragmentów dokumentów ww. klasami. Takie rozwiązanie ma z założenia szansę być bardziej odporne na wspomniane trudności. W szczególności pojawienie się zmiany po stronie pewnego portalu informacyjnego wymaga jedynie ponownego nauczenia klasyfikatora. Założeniem projektowym była też możliwość szybkiego tworzenie wielu klasyfikatorów — każdy klasyfikator przeznaczony dla innego portalu.

Habilitant samodzielnie opracował oryginalny algorytm i zaprogramował go w języku Java w ramach systemu pod nazwą „*Contents Extractor*". Program składał się z dwóch modułów: (1) modułu do przeprowadzania uczenia, oraz (2) modułu do ekstrakcji (rozpoznawania) treści z wykorzystaniem jednego z przygotowanych klasyfikatorów (przechowywanych w formie plików).

[21] Firma zarejestrowana w rejestrze przedsiębiorców prowadzonym przez Sąd Rejonowy dla m. st. Warszawy w Warszawie XIII Wydział Gospodarczy Krajowego Rejestru Sądowego pod numerem KRS: 0000321023, NIP: 951-227-09-31, REGON: 141677733.

W module do przeprowadzania nauki, po wskazaniu adresu URL dokumentu program w pierwszej kolejności dokonuje odpowiedniego rozkładu źródła HTML tego dokumentu na fragmenty. Następnie, program wylicza z nich pewne cechy przydatne na potrzeby klasyfikatora (więcej szczegółów algorytmicznych nie może być tu ujawnione ze względu na przekazane prawa autorskie firmie IVN). Poprzez przyrostowe uczenie, klasyfikator stopniowo coraz lepiej potrafi wychwytywać właściwą treść, dzięki czemu użytkownik-nauczyciel nie musi znakować kilkuset fragmentów HTML jako *tytuł*, *treść*, *nie-treść*, a zwykle znakuje zaledwie kilka (lub nawet zero) fragmentów. Praktyczne testy pokazały, że po zapoznaniu się z około 20–30 dokumentami HTML, klasyfikator bardzo rzadko wymagał dalszej interwencji nauczyciela. Całościowy czas przygotowania klasyfikatora dla pewnego portalu był przeciętnie krótszy niż 3 godziny (jest to więc nieporównywalnie lepsze, niż jakiekolwiek rozwiązanie wymagające zmian programistycznych).

Warto zwrócić uwagę, że w przedstawionym problemie mamy do czynienia z błędami dwojakiego rodzaju: (1) do wynikowego rozpoznania mogą wkradać się fragmenty nie będące prawdziwą treścią artykułu (*false positives*) oraz (2) niektóre fragmenty prawdziwej treści artykułu mogą zostająć błędnie rozpoznane jako nie-treść (*false negatives*). Oba rodzaje błędów są groźne i niepożądane. Błędy pierwszego rodzaju powodują, że osoba odsłuchująca słyszy fragmenty, nie mające sensu i związku z artykułem. Błędy drugiego rodzaju mogą skutkować brakiem pewnych istotnych informacji w artykule, a tym samym przekłamaniem jego treści. A zatem problem wymagał rozwiązania o wysokiej precyzji.

Po przeprowadzeniu obszernych testów, opracowany przez habilitanta **program wykazywał poprawność (zdolność do uogólniania) na poziomie powyżej 99%**. Poprawność mierzona była jako odsetek długości dobrze rozpoznanego tekstu w stosunku do długości całego tekstu podanego na wejście algorytmu (długość tekstu mierzona jako liczba znaków). Mając na uwadze błędy pierwszego i drugiego rodzaju dodatkowo testowano także **czułość i specyficzność — otrzymując również wyniki powyżej 99%**.

Jako załącznik nr 8 przedstawiono dokument od firmy Interactive Voice News potwierdzający stworzenie przez habilitanta i wdrożenie opisanego tu algorytmu oraz programu.

5d Informacje o osiągnięciach dydaktycznych

1. Opracowanie i prowadzenie wykładów w latach 2005–2011 (na Wydziale Informatyki ZUT, wcześniej Politechniki Szczecińskiej) z następujących przedmiotów:
    1. „Wprowadzenie do techniki",
    2. „Inżynierskie pakiety oprogramowania CAD/CAM",
    3. „Wstęp do sztucznej inteligencji",
    4. „Algorytmy eksploracji danych",
    5. „Metody rozpoznawania wzorców",
    6. „Projektowanie i programowanie systemów sztucznej inteligencji",
    7. „Metody sztucznej inteligencji w grach komputerowych",
    8. „Analiza danych i uczenie maszynowe",
    9. „Eksploracja wiedzy z internetowych repozytoriów danych".

2. Prowadzenie zajęć laboratoryjnych (wszystkie przedmioty wymienione w punkcie 1) oraz dodatkowo:

    1. „Metody sztucznej inteligencji",
    2. „Sieci neuronowe i aplikacje sztucznej inteligencji",
    3. „Zastosowania sztucznej inteligencji w technice, ekonomii i medycynie".

3. Prowadzenie prac inżynierskich i magisterskich studentów w latach 2005–2011. Łączna liczba prac (zakończonych pozytywną obroną): 27.

4. Opracowywanie materiałów dydaktycznych w wersji elektronicznej do wykładów i zajęć laboratoryjnych (dokumenty .pdf, skrypty MATLABa, programy w języku Java, zbiory danych do analizy). Materiały umieszczane są regularnie na stronie internetowej prowadzonej przez Zakład Sztucznej Inteligencji (Wydział Informatyki, ZUT): `http://wikizmsi.zut.edu.pl`.

5. Prowadzenie cotygodniowych zajęć sekcji brydża sportowego przy Klubie Uczelnianym AZS Szczecin od roku: 2008; w tym zdobycie ze studentami medali na Akademickich Mistrzostwach Polski (Wrocław, 2010 r.): II miejsce w turnieju par i III miejsce w punktacji generalnej uczelni technicznych.

6. Organizacja turniejów brydża sportowego dla studentów i pracowników z okazji Doby Sportu ZUT w trakcie juwenaliów. Lata: 2009, 2010, 2011.

7. Cykl indywidualnych lekcji ze „sztucznej inteligencji" dla zdolnego licealisty Krzysztofa Krzysztofika z Kamienia Pomorskiego (lata 2010-2011, lekcje finansowane przez liceum w Kamieniu Pomorskim).

5e Informacje o współpracy z instytucjami, organizacjami i towarzystwami naukowymi

1. Współpraca w latach 2007–2011 z Pomorską Akademią Medyczną w Szczecinie (obecnie: Pomorski Uniwersytet Medyczny) z profesorami: J. Lubińskim, A. Ciechanowiczem, a także z: dr M. Kaczmarczykiem i W. Piesiakiem. Praca nad problemami analizy zbiorów danych medycznych i wykrywania w nich reguł. W szczególności praca nad danymi genetycznych zbieranymi za pomocą mikromacierzy. Dane dotyczyły:

    - chorób nowotworowych,
    - choroby sodowrażliwości,
    - zagadnienia długowieczności.

2. Współpraca zagraniczna z Uniwersytetem w Modenie (Włochy) z prof. Gisellą Fachinetti przy projekcie badawczym p.t. *"Measuring Interaction Between Quality of Life, Children Well-Being, Work and Public Policies"*. Projekt zajmował się problemem socjologicznym: związkami pomiędzy jakością życia, dzieciństwem i warunkami prawnymi w różnych krajach. W ramach tej współpracy habilitant pracował w szczególności nad: ankietyzacją młodych ludzi w Polsce oraz programem do wykrywania reguł decyzyjnych na podstawie danych.

3. Wykonanie recenzji artykułów dla:
   - czasopisma *International Journal of Neural Systems* (INT J NEURAL SYST, ISSN: 0129-0657),
   - czasopisma *Pattern Recognition Letters* (PATTERN RECOGN LETT, ISSN: 0167-8655),
   - czasopisma *Metody Informatyki Stosowanej* (ISSN: 1898-5297),
   - konferencji międzynarodowej International Conference on Artificial Intelligence and Soft Computing ICAISC, Zakopane,
   - konferencji międzynarodowej Advanced Computer Systems (organizowanej przez Wydział Informatyki, ZUT).
4. Współpraca z grupą studencką *Netcamp* — przeprowadzenie analizy zbioru danych na temat umiejętności studentów Wydziału Informatyki, sierpień/wrzesień 2011 r.

5f Informacje o działalności popularyzującej naukę

1. Wygłoszenie referatu p.t. „Sztuczna inteligencja w grach komputerowych" dla uczniów szkół średnich w ramach X Zachodniopomorskiego Festiwalu Nauki, Szczecin, 2010 r.
2. Zorganizowanie otwartego seminarium p.t. „Wykrywanie dobrych heurystyk do gry w warcaby" z udziałem dyplomanta Mateusza Bożykowskiego — laureata konkursów programistycznych. W trakcie seminarium miała miejsce prezentacja programu warcabowego opracowanego przez dyplomanta pod kierunkiem dra Klęska oraz **rozegranie meczu pokazowego z zawodnikiem warcabistą**, maj 2009 r.
3. Udział w dwóch audycjach radiowych pt. „Pulsar" popularyzujących naukę w latach 2007 i 2011. Audycja opracowywana przez Polskie Radio Szczecin (redaktor: Marek Borowiec).
4. Organizowanie spotkań otwartego seminarium Zakładu Sztucznej Inteligencji (WI, ZUT) — referaty naukowe na temat aktualnie prowadzonych badań i dowolnych problemów matematycznych, wzajemne szkolenie się pracowników zakładu. Spotkania (zwykle cotygodniowe) w latach 2004–2011.

5g Publiczna prezentacja

Poniższy spis przedstawia wybrane ważniejsze referaty publiczne habilitanta.

1. Wystąpienie na międzynarodowej konferencji ICAART 2011 w Rzymie. (30 min., full paper). Artykuł: "Relationship between cross-validation and Vapnik bounds on generalization of learning machines", 2011 r.
2. Wystąpienie na seminarium Polskiej Akademii Nauk (oddział Gdańsk) w Szczecinie. Referat: „Ukryte Modele (Łańcuchy) Markowa — algorytmy i zastosowania", 2009 r.
3. Sesja plakatowa na międzynarodowej konferencji ICAISC 2008 w Zakopanem. Artykuł: "Maximal Margin Estimation with Perceptron-Like Algorithm", 2008 r.
4. Wystąpienie na międzynarodowej konferencji ACS w Międzyzdrojach. Artykuł: "Mining Interesting Rules and Patterns for Salt-Sensitivity of Blood Pressure", 2008 r.

5. Sesja plakatowa na międzynarodowej konferencji ICAISC 2006 w Zakopanem. Artykuł: "Construction of a Neurofuzzy Network Capable of Extrapolating (and Interpolating) with Respect to the Convex Hull of a Set of Input Samples in $\mathbb{R}^n$", 2006 r.

6. Wystąpienie na międzynarodowej konferencji ACS w Ełku. Artykuł: "The Jeep Problem, searching for the best strategy with a genetic algorithm", 2004 r.

7. Wystąpienie na seminarium *Warsaw International Seminar on Intelligent Systems* w Warszawie. Referat: "On the need for extrapolation in input-output modeling", 2004 r.

8. Wystąpienie na międzynarodowej konferencji ACS w Międzyzdrojach. Artykuł: "Control over m-th order extrapolation capabilities of neuro-fuzzy models of multidimensional systems", 2003 r.

9. Wystąpienie na międzynarodowej konferencji ACS w Międzyzdrojach. Artykuł: "Algorithm for Automatic Definition of Validated and Non-Validated Region in Multi-Dimensional Space", 2003 r.