DOBÓR CECH DIAGNOSTYCZNYCH DLA KLASYFIKATORA SVM W ZADANIU KLASYFIKACJI STANU PRZETOKI TĘTNICZO-ŻYLNEJ NA PODSTAWIE SYGNAŁU AKUSTYCZNEGO

THE SELECTION OF FEATURES FOR THE SVM CLASSIFIER IN THE ARTERIOVENOUS FISTULA STATE ESTIMATION ON THE BASIS OF ACOUSTIC SIGNAL

Marcin Grochowina*, Lucyna Leniowska

Uniwersytet Rzeszowski, Wydział Matematyczno-Przyrodniczy,
Katedra Mechatroniki i Automatyki, 35-959 Rzeszów, al. Rejtana 16c

* e-mail: gromar@ur.edu.pl

STRESZCZENIE

W artykule przedstawiono proces selekcji cech diagnostycznych dla klasyfikatora SVM. Badania przeprowadzone zostały z użyciem zbioru danych zawierającego próbki sygnału dźwiękowego emitowanego przez przetokę tętniczo-żylną. Celem prac było stworzenie rozwiązania klasyfikacji wieloklasowej w oparciu o klasyfikator z rodziny SVM pozwalającego na skuteczną i wiarygodną ocenę stanu przetoki tętniczo-żylnej.

Słowa kluczowe: klasyfikator SVM, przetoka tętniczo-żylna, selekcja cech

ABSTRACT

The paper presents the process of selection diagnostic features for SVM classifier. The study was conducted with using a data set containing samples of the sound signal emitted by the arteriovenous fistula. The objective was to create a solution multi-class classification based on SVM classifier family allowing for an effective and reliable evaluation of the arteriovenous fistula state.

Keywords: SVM classifier, arteriovenous fistula, features selection

1. Wstęp

Każdy proces klasyfikacji opiera się na zestawie cech dostarczanych do klasyfikatora na podstawie których zostaje podjęta decyzja i zwrócony wynik, a właściwy dobór zestawu cech istotnie wpływa na podniesienie jakości procesu klasyfikacji. Rozwiązaniem zapewniającym najlepszą jakość jest przetestowanie wszystkich możliwych niepustych podzbiorów zbioru wejściowego. Niestety liczba niepustych podzbiorów zbioru *n*-elementowego wynosi 2^n*-1*. Implikuje to możliwość zastosowania pełnego przeglądu podzbiorów jedynie dla n nieprzekraczającego kilkunastu elementów, gdyż pełna analiza dla większych wartości *n* jest zbyt czasochłonna. Konieczne jest zatem zastosowanie metod quasi-optymalnych pozwalających wyłonić podzbiór cech o możliwie dużej sprawności klasyfikacyjnej [1].

Klasyfikator SVM (ang. *Support Vector Machine*) jest uznawany za wysokiej klasy rozwiązanie w dziedzinie klasyfikacji [2]. Generowane przez algorytm SVM reguły decyzyjne umożliwiają zwykle osiągnięcie dobrej jakości klasyfikacji i z tego względu SVM jest często stosowany w badaniach z różnych dziedzin [3]. Choć klasyfikator SVM nie jest wrażliwy na nadmiarowe, redundantne lub szumowe dane tak jak np. klasyfikatory z rodziny najbliższego sąsiedztwa, ograniczenie ilości cech wejściowych do optymalnego minimum korzystnie wpływa na jakość klasyfikacji. Ponadto, ponieważ klasyfikator SVM w procesie treningu buduje globalny model aproksymacji badanego zjawiska w postaci hiperpłaszczyzny w *n*-wymiarowej przestrzeni cech, redukcja ilości cech w wejściowym zestawie korzystnie wpływa na zmniejszenie stopnia skomplikowania uzyskanego modelu.

2. Dane

W badaniach użyto zbioru danych uzyskanych w efekcie przetworzenia dźwięku emitowanego przez przetokę tętniczo-żylną. Przeprowadzone dotychczas badania wykazały, że charakter dźwięku emitowanego przez krew przepływającą przez przetokę tętniczo-żylną jest zależny od stanu przetoki [4, 5, 6].

Pozyskano łącznie 1190 próbek pochodzących od 19 pacjentów posiadających przetokę zlokalizowaną na nadgarstku. Pobranie materiału polegało na rejestracji dźwięku emitowanego przez krew przepływającą przez przetokę tętniczo-żylną. Materiał pobrany został dedykowaną głowicą pomiarową wyposażoną w mikrofon elektretowy CZ034 produkcji Ringford, o czułości -42 dB (0 dB = 1 V/Pa, 1 kHz), tj. 8 mV/Pa i odstępie sygnał/szum niemniejszym niż 60 dB. Do rejestracji sygnału użyto zintegrowanej karty dźwiękowej RV730 będącej częścią układu Radeon 4000 produkcji AMD oraz dedykowanego oprogramowania działającego pod kontrolą systemu operacyjnego Linux. Szybkość próbkowania sygnału ustalono na 8000 próbek/s. Zarejestrowany materiał poddano obróbce numerycznej z użyciem pakietów Matlab oraz Octave celem ekstrakcji cech charakteryzujących sygnał.

Z pobranego materiału wyodrębniono 23 cechy; 6 w dziedzinie czasu i 18 w dziedzinie częstotliwości. Cechy w dziedzinie czasu: t0, t4, y0, y4, p0 oraz p4 opisują parametry czasowe, amplitudowe oraz kształt obwiedni sygnału w obrębie pojedynczego okresu rytmu serca. Cechy w dziedzinie częstotliwości: f1-f17 opisują gęstość widma częstotliwości zarejestrowanego sygnału w określonych przedziałach z zakresu 20−600 Hz. Cecha fmax oznacza częstotliwość dla której moduł widma częstotliwości osiągną wartość największą.

Przetoki pacjentów ocenione zostały jako sprawne jednakże w niejednakowym stopniu. Wyodrębniono 8 grup reprezentujących przetoki o różnym stopniu nasilenia patologii. Grupy zostały oznaczone literowymi etykietami ***a***-***h***, przy czym w grupie ***a*** znalazły się przetoki w najlepszym stanie natomiast w grupie ***h*** w stanie najgorszym.

Klasyfikację zgromadzonych danych przeprowadzono przy użyciu pakietu WEKA 3.7.13 uruchomionym z użyciem JRE Oracle Java 1.8 [7]. Obliczenia wykonywane były na komputerze z procesorem Intel Core2 T6570 2.1GHz. Podczas pomiarów zapotrzebowania czasowego algorytmów używano jedynie jednego rdzenia procesora.

3. Metody

W ramach badań przetestowanych zostało pięć metod selekcji cech. Cztery pierwsze są metodami powszechnie znanymi i dostępnymi w pakiecie WEKA [8]. Metoda piąta jest metodą autorską opracowaną na potrzeby tego konkretnego zadania.

Użyte metody to:

- Correlation − buduje ranking cech oceniając każdą z nich z osobna. Kryterium oceny jest wartość bezwzględna współczynnika korelacji cechy z klasą. Im wyższy współczynnik korelacji, tym wyższe położenie cechy w rankingu.
- SVMeval − ocenia jakość cech używając liniowej sieci SVM. Jakość cechy określana jest na podstawie wagi przypisanej jej w liniowym klasyfikatorze SVM.
- PCA − dokonuje liniowej transformacji cech w inną przestrzeń, w której cechy wchodzące w skład nowego zestawu są ze sobą nieskorelowane i posortowane pod względem ilości informacji wnoszonej w proces klasyfikacji

- Forward search − metoda budująca zestaw cech poczynając od jednej i dodając sukcesywnie te cechy, które zapewniają najlepszą jakość klasyfikatora. Metoda ta oparta jest o klasyfikator, który ma zostać użyty w docelowym rozwiązaniu, w tym przypadku − SVM.
- Joined pairs [9] − metoda autorska polegająca na budowie rankingu par cech. Pary oceniane są na podstawie jakości zbudowanego przy ich użyciu klasyfikatora. Następnie budowany jest ranking cech. Do listy dodawane są pary cech w kolejności od najlepszej. Do listy dodawane są jedynie te cechy, które nie występują jeszcze na liście.

W badaniach użyty został klasyfikator SVM z radialną funkcją jądra.

Ocena jakości klasyfikacji oparta została na wskaźniku F-measure. Wartość tego wskaźnika może przyjmować wartości z przedziału ⟨0; 1⟩ i jakość klasyfikacji jest tym wyższa im wartość F-measure jest bliższa 1. Podział zbioru na część uczącą i testową przeprowadzony został z użyciem dziesięciokrotnej kroswalidacji.

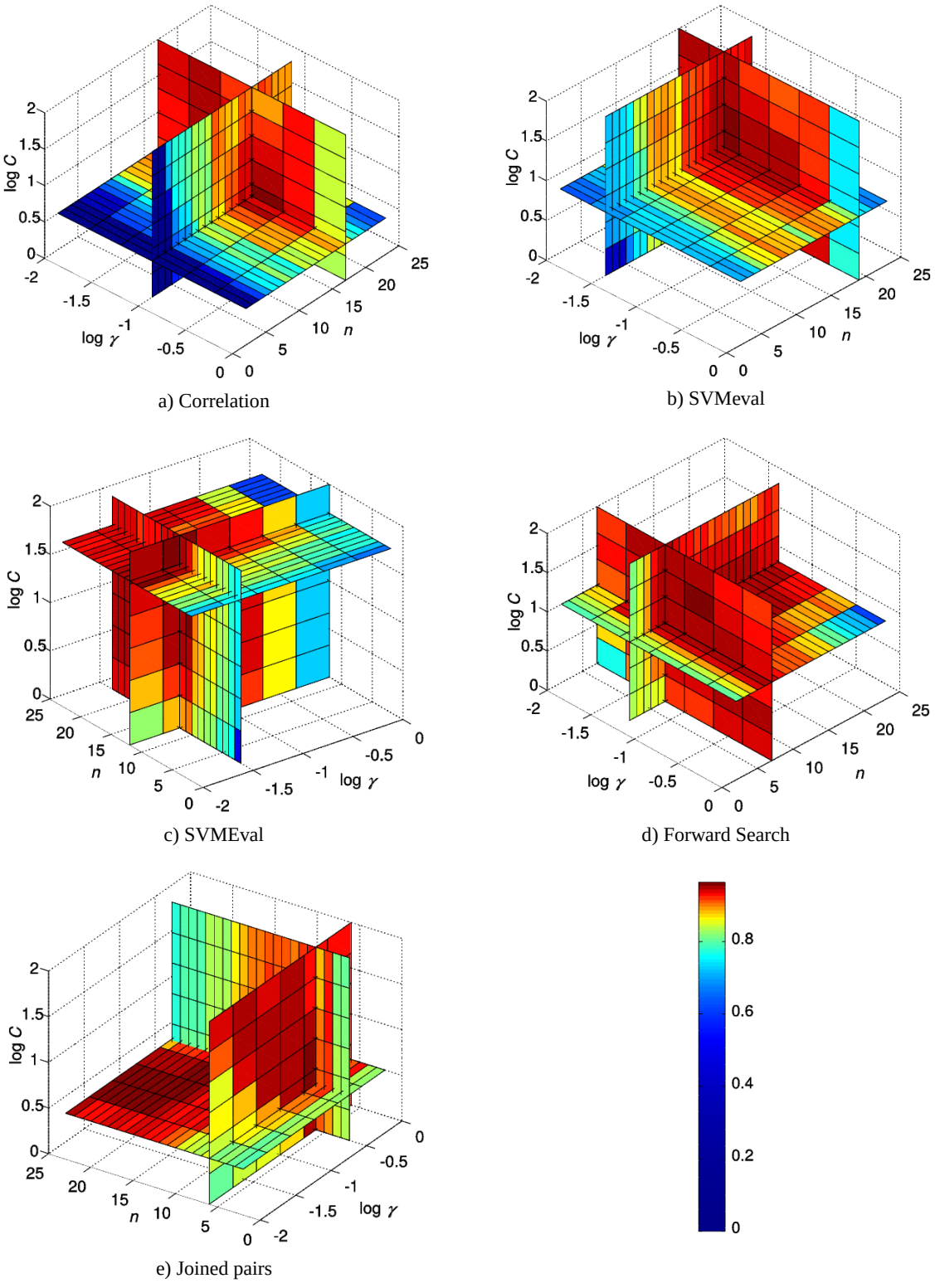
4. Wyniki

Jakość każdego zestawu cech otrzymanych za pomocą każdej z metod oceniana była na podstawie jakości klasyfikatora zbudowanego na podstawie tego zestawu. Dla każdego zestawu cech wygenerowano 22 podzbiory zawierające kolejno od 2 do 23 cech. W kolejnych podzbiorach cechy były dołączane w kolejności wskazanej rankingiem.

Tabela 1. Ranking cech

Lp.	Correlation	SVMeval	PCA	Forward search	Joined pairs
1	f14	f11	v1	f11	
2	f15	f5	v2	f3	f11,f3
3	f8	f14	v3	f1	f12
4	f16	f16	v4	f13	f13
5	f7	f13	v5	f4	f1
6	f6	f9	v6	f14	f4
7	f13	f15	v7	f10	f5
8	f9	f3	v8	f16	f9
9	f5	f8	v9	f8	f2
10	f4	f12	v10	f15	f10
11	f10	f6	v11	f12	f7
12	f12	f7	v12	y4	y0
13	f11	f10	v13	f9	f8
14	f3	f1	v14	t4	y4
15	f1	f2	v15	f2	t4
16	f2	f4	v16	fm	fm
17	fm	fm	v17	f6	f6
18	t1	y4	v18	f5	f16
19	y0	t4	v19	y0	t1
20	p04	y0	v20	p01	f14
21	t4	t1	v21	f7	f15
22	y4	p01	v22	p04	p01
23	p01	p04	v23	t1	p04

Wstępnie dokonano oszacowania wartości granicznych dla parametrów pracy klasyfikatora SVM, tj. C oraz γ, a następnie zbudowano i przetestowano klasyfikatory dla określonej liczby wartości tych parametrów z wnętrza oszacowanych przedziałów. Ponieważ wpływ zmian parametrów C oraz γ na pracę klasyfikatora SVM jest silnie nieliniowy, przyjęto wewnątrz przedziałów skalę logarytmiczną



a) Correlation

b) SVMeval

c) SVMEval

d) Forward Search

e) Joined pairs

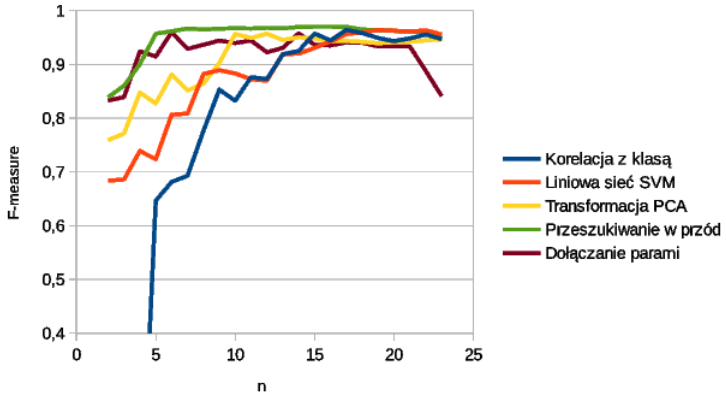
Rys. 1. F−measure = f(C,γ,n)

Dla każdego podzbioru badano klasyfikator dla parametrów C = {1; 3; 6; 10; 30; 60; 100} oraz γ = {0,01; 0,03; 0,06; 0,1; 0,3; 0,6; 1} w celu wyłonienia optymalnych parametrów pracy klasyfikatora. Graficzne zestawienie wyników obliczeń przedstawione zostało na rysunku 1.

Każda z płaszczyzn wykresu prezentuje wartość wskaźnika F-measure dla optymalnej wartości

parametrów C, γ oraz liczby cech w zestawie. Punkt przecięcia płaszczyzn wykresu wyznacza maksymalną uzyskaną jakość klasyfikacji.

Na rysunku 2 przedstawiono zależność jakości klasyfikacji w funkcji liczby cech w zestawie dla wartości C oraz γ optymalnych dla danego algorytmu selekcji cech.



Rys. 2. F−measure = f(n) dla optymalnych wartości C i γ

Tabelaryczne zestawienie wyników uzyskanych dla klasyfikatora SVM zaprezentowano w tabeli 2.

Tabela 2. Zbiorcze zestawienie wyników dla klasyfikatora SVM

	Correlation	SVMeval	PCA	Forward search	Joined pairs
czas budowy rankingu	00:00:01	00:00:10	00:00:01	47:20:15	31:05:44
czas oceny zestawu	06:46:04	06:46:04	00:00:00	00:00:00	06:46:04
łączny czas	06:46:05	06:46:14	00:00:01	47:20:15	37:51:48
optymalne n	17	19	12	7	6
optymalne γ	0,1	0,03	0,03	0,06	0,3
optymalne C	3	6	60	10	3
dokładność (F-measure)	0,96	0,96	0,96	0,97	0,96

5. Dyskusja

Metody Joined pairs oraz Forward search zapewniły uzyskanie najwyższej jakości klasyfikacji przy niewielkiej liczbie cech wejściowych. Zestawy cech wygenerowane algorytmem Forward search wykazują największą stabilność klasyfikacji, tj. utrzymanie maksymalnej jakości klasyfikacji w trakcie dodawania kolejnych cech do zestawu. Algorytm Joined pairs pozwolił na uzyskanie równie szybkiej zbieżności do wartości maksymalnej F−measure, jednak nie zapewnił równie dobrej jak Forward search stabilności rozwiązania.

Maksymalna jakość klasyfikacji dla każdego z algorytmów selekcji cech okazała się porównywalna, jednakże liczba cech niezbędnych do uzyskania maksimum F−measure znacząco się różniła. Najlepsze okazały się metody Joined pairs oraz Forward search osiągając najlepszą jakość klasyfikacji dla liczby cech 2−3-krotnie mniejszej niż pozostałe metody.

Wysoka jakość klasyfikacji oceniona za pomocą wskaźnika F-measure na poziomie powyżej 0,95 wydaje się być zbyt optymistyczna. Może to być wynikiem zastosowania w procesie testowania kros-walidacji, która spowodowała, że zarówno w zbiorze uczącym jak i testowym znajdowały się dane pochodzące od tych samych pacjentów, a więc potencjalnie silnie ze sobą skorelowane.

W celu weryfikacji uzyskanych wyników konieczne jest przeprowadzenie dodatkowych testów z zapewnieniem użycia danych wzajemnie nieskorelowanych, to jest wyłączając kolejno w charakterze zbioru testowego dane pochodzące od pojedynczych pacjentów. Dodatkowo korzystne dla jakości uzyskanych wyników będzie zwiększenie liczności zbioru danych i poszerzenie go o próbki pobrane od większej liczby pacjentów.

LITERATURA

[1] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, V. Vapnik: *Feature selection for SVMs*, Neural Information Processing Systems Foundation, 2000.

[2] C. Cortes, V. Vapnik: *Support-vector networks*, Machine learning, vol. 20(3), 1995, s. 273−297.

[3] R.E. Fan, R.H. Chen, C.J. Lin: *Working set selection using second order information for training support vector machines*, The Journal of Machine Learning Research, vol. 6, JMLR org, 2005, s. 1889–1918.

[4] M. Grochowina, L. Leniowska, P. Dulkiewicz: *Application of Artificial Neural Networks for the Diagnosis of the Condition of the Arterio-venous Fistula on the Basis of Acoustic Signals*, Brain Informatics and Health, Springer, 2014, s. 400–411.

[5] M. Grochowina, L. Leniowska, P. Dulkiewicz: *Comparison of SVM and k-NN classifiers in the estimation of the state of the arteriovenous fistula problem*, Computer Science and Information Systems (FedCSIS), 2015 Federated Conference on, IEEE, 2015, s. 249–254.

[6] M. Grama, J. Tranholm Olesena, H.Ch. Riisa, M. Selvaratnama, M. Urbaniaka: *Stenosis detection algorithm for screening of arteriovenous fistulae*, 15th Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC 2011), Springer, 2011, s. 241–244.

[7] "WEKA documentation," http://www.cs.waikato.ac.nz/ml/weka/documentation.html

[8] R.R. Bouckaert, E. Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald, D. Scuse: "WEKA Manual," University of Waikato, 2013.

[9] M. Grochowina, L. Leniowska: *The new method of the selection of features for the k-NN classifier in the arteriovenous fistula state estimation*, Computer Science and Information Systems (FedCSIS), 2016 Federated Conference on. IEEE, 2016.

otrzymano / submitted: 22.11.2016
wersja poprawiona / revised version: 27.11.2016
zaakceptowano / accepted: 19.12.2016