Agnieszka Duraj
Instytut Informatyki, Politechnika Łódzka
Wólczańska 215, 90-924 Łódź
email: agnieszka.duraj@p.lodz.pl

WYKRYWANIE WYJĄTKÓW PRZY UŻYCIU WEKTORÓW NOŚNYCH

Streszczenie – W artykule omówiono metodę wektorów nośnych w bardzo ważnym aspekcie wykrywania wyjątków w dużych zbiorach danych. Wykrywanie wyjątków jest istotne przy procesach analizowania danych, gdzie mamy do czynienia z klasyfikacja, grupowaniem, wyznaczaniem reguł asocjacyjnych. Istnienie wyjątków w klasyfikowanych czy grupowanych danych wpływa na cały kontekst analizy. Może prowadzić do wyznaczenia błędnych reguł. Wyjątki muszą być zatem identyfikowane. W niniejszym artykule skupiono się na metodzie wektorów nośnych w celu wykrywania wyjątków. Badania zostały oparte na zbiorach z repozytorium UCI [17].

Słowa kluczowe: wykrywanie wyjątków, maszyna wektorów nośnych

1 Wprowadzenie

Eksploracja danych (ang. data mining) jako etap procesu odkrywania wiedzy z baz danych (ang. Knowledge Discovery in Databases, KDD) stanowi obecnie bardzo popularny dział badań naukowych. Ideą tej dziedziny jest szybkie znajdowanie ukrytych dla ludzkiego oka prawidłowości lub tez nieprawidłowości w zbiorze danych. Jako proces analityczny zajmuje się przetwarzaniem bardzo dużych zasobów w celu odnalezienia regularnych wzorców, współzależności między zmiennymi. Można zatem powiedzieć, że przy wykorzystaniu szybkości procesorów za pomocą algorytmów data mining odkrywamy ukryte dla człowieka prawidłowości w danych. Prowadzimy do zdefiniowania i przewidywania wielkości sprzedaży, zachowań i profilu klienta. Cały proces eksploracji danych podzielony jest na etapy:

- wstępnego przetwarzania danych,
- budowania odpowiedniego modelu, jego oceny i weryfikacji działania,
- zastosowania modelu dla nowych obiektów.

Wstępne przetwarzanie danych prowadzi do czyszczenia zbioru z rekordów z pustych atrybutów i tzw. brudnych danych. Na tym etapie określa się również najważniejsze cechy w kontekście prowadzonej analizy. Ma to ogromne znaczenia w procesach klasyfikacji, czy też

grupowania. Wyjątki w zbiorach danych traktowane jako brudne dane powstające w wyniku uszkodzenia systemu, błędu człowieka należy oczywiście usunąć z bazy i pominąć w trakcie analizy. Jednak na tym etapie nie możemy wyeliminować tzw. wyjątków, które mogą zmieniać kontekst analizy ale są szczególnymi przypadkami. Wykrywanie wyjątków jest zatem bardzo istotnym zagadnieniem w eksploracji danych. Metody wykrywające wyjątki są różnorodne. Oparte są przede wszystkim na metodach statystycznych, miarach odległości czy też funkcjach podobieństwa i niepodobieństwa. Bogaty przegląd tej dziedziny podano w pracach [1,2]. Przegląd metod wykrywania dla danych medycznych podano zaś w [6,9]. Wykrywanie wyjątków przy użyciu podsumowań lingwistycznych zaproponowano w pracach Duraj i współautorzy [7,8,9]. Innowacyjne metody dotyczące wykrywania wyjątków związane z algorytmami genetycznymi podano w [4,5]. Metodę wykrywania wyjątków określających anomalie między zadaniami zapobiegającymi występowaniu konfliktów zasobów w przygotowanym harmonogramie zaproponował Smoliński w pracach [13-15]. Z kolei w [12] podano metodę wykrywania wyjątków spowodowanych określonym zjawiskiem fizycznym.

W niniejszym artykule skupiamy się na wykrywaniu wyjątków przy użyciu bardzo popularnego i dobrego klasyfikatora jakim jest maszyna wektorów nośnych. Metoda wektorów nośnych (ang. Support Vector Machines - SVM) wprowadzona przez Vladimir N. Vapnik'a [19] tworzy przestrzenie decyzyjne. Przestrzenie te wyznacza dzieląc całą przestrzeń według tworzonych granic separujących obiekty. W najprostszej postaci dzieli przestrzeń na dwie podprzestrzenie – dwie klasy i oddziela je linią graniczną. Obiekt nieznany w zależności w której przestrzeni się znajdzie do tej przestrzeni zostanie zaklasyfikowany. Ten najprostszy przypadek z dwoma hiperpłaszczyznami staje się intuicyjny i prosty. Zbliżony do zagadnienia regresji liniowej . Istnieje kilka typów wektorów nośnych, z różnymi funkcjami bazowymi. Są to na przykład: liniową, wielomianową funkcja bazowa, RBF (radialne funkcje bazowe), sigmoidalna funkcja bazowa, Gaussowska funkcja bazowa, itd. W badaniach własnych skupiono się na gaussowskiej funkcji bazowej.

Wykrywanie wyjątków stanowi bardzo ważny aspekt eksploracji danych. Poszukiwanie nowych metod lub modyfikacja istniejących algorytmów jest jak najbardziej uzasadniona. Układ niniejszej pracy jest następujący: w sekcji 2 omawiamy podstawową ideę metody wektorów nośnych. Następnie podajemy kontekst wykrywania wyjątków w oparciu o obiekt lokalny i globalny. Sekcja 4 to wyniki z przeprowadzonych eksperymentów badawczych. Praca zakończona jest wnioskami.

2 Idea działania maszyny wektorów nośnych

Algorytm wektorów nośnych (ang. Support Vector Machine) (SVM) wprowadzony przez Vladimir N. Vapnik'a [19] jest często używany do klasyfikacji czy też predykcji danych. Polega on głownie na wybraniu najlepszej z użytych hiperpłaszczyzn dyskryminacyjnych. Istotna jest zatem maksymalizacja marginesu separacji pomiędzy dwoma klasami, przy zachowaniu najmniejszego błędu klasyfikacji. SVM jest używany w wielu różnych dziedzinach począwszy od analizowania danych po rozpoznawanie mowy, tekstu czy klasyfikacji, analizę danych finansowych, medycznych. Główną zaletą maszyny wektorów nośnych staje się możliwość przetwarzania danych nienumerycznych, strumieni danych. Istotny jest w tym przypadku odpowiedni dobór funkcji jądra oraz konstrukcja i odnalezienie hiperpłaszczyzny separującej punkty należące do dwóch lub wielu klas. Margines pomiędzy dwoma zbiorami danych powinien być wyznaczany jako maksymalny (największy) [18,19].

Opis sposobu tworzenia hiperpłaszczyzn można omówić najprościej dla modelu liniowego. Model nieliniowy nie jest używany w badaniach w związku z tym nie jest omawiany w tym punkcie artykułu.

Model liniowy

Niech x_i będzie wektorem wejściowym zaś y_i etykietą klasy przyjmującą wartości {-1,1}. Rozpatrywany jest zbiór uczący jako para (x_i, y_i) dla i=1,2…,p, $x_i \in R^d$.

Załóżmy, że klasy y_i są liniowo separowane. Wówczas funkcja $g(\mathbf{x})$ zdefiniowana równaniem (1) będzie hiperpłaszczyzną rozdzielającą obie klasy.

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0 \tag{1}$$

gdzie $\mathbf{w} = [w_1, w_2, w_3, ..., w_N]^T$, $\mathbf{x} = [x_1, x_2, x_3, ..., x_N]^T$.

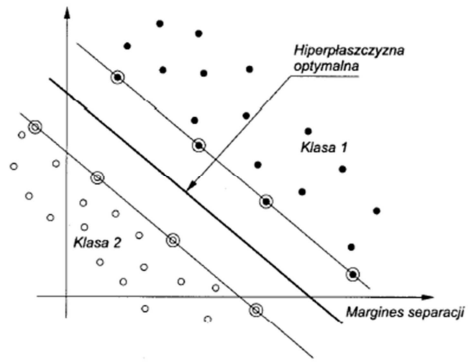
Jeżeli spełnione są założenia (2) optymalną hiperpłaszczyznę, która maksymalizuje margines separacji możemy zapisać równaniem (3) zaś odległość *odl* wybranego obiektu $\mathbf{x}$ od optymalnej hiperpłaszczyzny równaniem (4).

$$\begin{cases} \mathbf{w}^T\mathbf{x} + b > 0 & dla \quad y_i = 1 \\ \mathbf{w}^T\mathbf{x} + b < 0 & dla \quad y_i = -1 \end{cases} \tag{2}$$

$$g(\mathbf{x}) = \mathbf{w}_0^T \mathbf{x} + b_0 = 0 \tag{3}$$

$$odl(\mathbf{x}) = \frac{g(\mathbf{x})}{\|\mathbf{w}_0\|} \tag{4}$$

Interpretację graficzną tworzonych hiperpłaszczyzn metodą wektorów nośnych pokazano na Rys. 1.



Rys. 1. Wizualizacja maszyny wektorów nośnych Źródło: [16]

Punkt leżący najbliżej optymalnej hiperpłaszczyzny tworzy para (x_i, y_i) dla której $\mathbf{w}^T\mathbf{x} + b = 1$ dla $y_i = 1$ oraz $y_i = -1$. Margines separacji wyznaczamy zgodnie z równaniem (5)

$$\rho = \frac{2}{\|\mathbf{w}_0\|} \tag{5}$$

Rozwiązanie zagadnienia maksymalizacji ρ jest równoznaczne z minimalizacją normy euklidesowej wektora wag $\|\mathbf{w}\|$ przy warunku $\min\{\mathbf{w}^T\mathbf{w}/2\}$ oraz ograniczeniach $y_i(\mathbf{w}^T\mathbf{x}+b) \geq 1$. W celu rozwiązania problemu optymalizacji stosuje się mnożniki Lagrange'a. Następuje wówczas minimalizacja prymarnej funkcji Lagrangea Lp lub maksymalizacja dualnej funkcji Lagrange'a Ld, opisanymi odpowiednio przez równania (6) oraz (7).

$$L_p(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{p} \alpha_i y_i (x_i w_i + b) + \sum_{i=1}^{p} \alpha_i \tag{7}$$

$$L_D = \sum_{i=1}^{p} \alpha_i - \frac{1}{2} \sum_{i=1}^{p} \sum_{j=1}^{p} \alpha_i \alpha_j y_i y_j x_i x_j \tag{8}$$

gdzie $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_p]^T$ jest wektorem mnożników Lagrange'a.

Równanie (7) dla danych nieseparowalnych liniowo można zapisać w postaci równania (9) gdzie ξ_i oznacza nieujemną zmienną dopełniającą, zaś φ oznacza wagę wybraną przez użytkownika określającą traktowanie błędów testowania w stosunku do wyznaczonego marginesu.

$$\min(\frac{1}{2}\mathbf{w}^T\mathbf{w} + \varphi\sum_{i=1}^{p}\xi_i) \tag{9}$$

Dla $\xi_i \geq 0$ otrzymujemy ograniczenie (10)

$$y_i(\mathbf{w}^T\mathbf{x} + b) \geq 1 - \xi_i \tag{10}$$

Maksymalne górne oszacowanie określone jest jako $\sum_{i=1}^{p}\xi_i$ jest granicą, maksymalnym górnym ich oszacowaniem. Dla funkcji L_D ograniczenie (10) zamienia się w (11).

$$0 \leq \alpha_i \leq \varphi \tag{11}$$

Niezerowe wartości mnożników Lagrange'a α_i z funkcjami ograniczeń równymi zeru, oznaczone jako M_v tworzą dla L_D zadania dualnego optymalne wagi hiperpłaszczyzny w postaci (12)

$$qaw_0 = \sum_{i=1}^{M_v}\alpha_i y_i x_i \tag{12}$$

Poddając dane odpowiednim transformacjom istnieje możliwość zastosowania metody wektorów nośnych dla wzorców nieseparowalnych liniowo, ten przypadek w niniejszej pracy nie jest omawiany.

3 Wyjątki w zbiorach danych

W analizie danych wyjątki to obiekty, które w znaczący sposób różnią się od pozostałych obiektów w zbiorze danych. Mogą wynikać z błędów użytkownika, błędów aparatury pomiarowej. Wówczas już na etapie wstępnego przetwarzania mogą być odnalezione i usunięte aby nie wpływać niekorzystnie na dalszy proces analizy. Mogą również opisywać obiekt, dla którego cechy tego obiektu znacznie różnią się od pozostałych obiektów. Wówczas analiza danych powinna być wykonana na zbiorze zawierającym takie wyjątkowe obiekty bowiem w takim przypadku obiekt i jego cechy reprezentuje niecodzienne zachowanie systemu.

W badaniach naukowych występuje wiele metod wykrywających wyjątki. Są one różnorodne ze względu na sposób działania samego

algorytmu, jak i stosowanego typu danych. Rozróżnia się metody oparte na podejściu statystycznym, prawdopodobieństwie, odległości, podobieństwie.

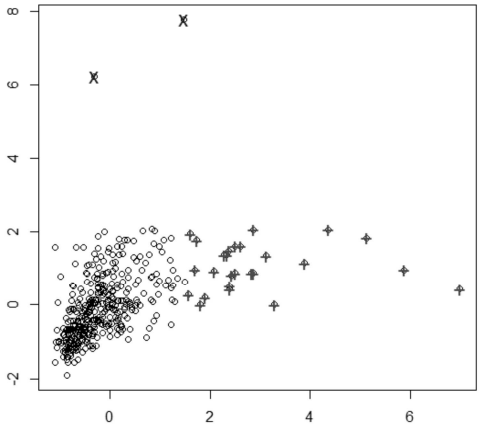
Wykrywanie wyjątków w oparciu o miarę odległości (ang. distance-basedoutliers) wprowadza pojęcie obiektu lokalnego oraz obiektu globalnego. Obiekt w zbiorze danych jest wyjątkiem odległościowym (globalnym) wtedy i tylko wtedy, gdy odległość co najwyżej k obiektów tego zbioru od analizowanego obiektu jest mniejsza od zadanej odleglości dist, wprowadzonej przez użytkownika. Wartość dist powinna być dobrana bardzo starannie. Spełnione musi być równanie (13), gdzie O, O' oznacza obiekty, zas d(O,O') jest miarą odległości między tymi obiektami, p zaś progiem ustalanym przez użytkownika.

$$\frac{\left|\{O' \mid d(O,O') \le dist\}\right|}{|D|} \le p \tag{13}$$

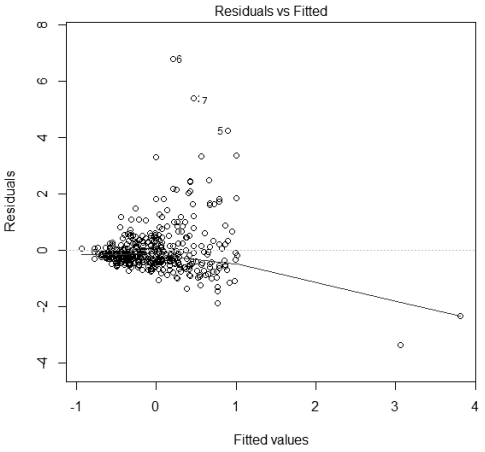
Problem z wykryciem punktów osobliwych może pojawić się w przypadku przestrzeni o dużej liczbie wymiarów, gdyż wszystkie znajdujące się w niej obiekty są w podobnej odległości od siebie. W celu wykrycia rzeczywistych punktów osobliwych konieczne jest bardzo staranne dobranie wartości parametru dist. W badaniach często używany jest algorytm wykrywający lokalne wyjątki w oparciu o tzw. lokalny współczynnik wyjątkowości (ang. Local outlier factor - LOF). Współczynnik ten określa jak wysoki jest stopień wyjątkowości danego obiektu. Zobacz szerzej w [1,3,10,11,].

4 Badania eksperymentalne

W badaniach użyto język R i pakiet kernlab bezpośrednio związany z klasyfikacją danych metodą wektorów nośnych. Wykrywanie wyjątków testowano na zbiorach danych pochodzących z repozytorium Machine Learning [17]. Zbiory zostały przeanalizowane również innymi klasyfikatorami. Użyto modelu regresji, klasyfikatora Bayesa. Wykonano także wykres diagnostyczny Rys. 2. Na Rys. 3 przedstawiono zaś zależność między modelem regresji a wartościami reszt. Łatwo zauważyć obiekty będące wyjątkami. Dodatkowo oznaczono je etykietami 5,6,7.



Rys. 2. Ilustracja graficzna przedstawiająca rozrzut danych – wykres diagnostyczny



Rys. 3. Graficzne przedstawienie zależności między przewidywanymi wartościami a resztą w modelu regresji z zaznaczonymi wyjątkami

W procesie uczenia klasyfikatora metodą wektorów nośnych użyto jako funkcję jądra funkcję Gaussowską. Dokładność dopasowania ustalono na poziomie 0.05. Ustalono również eksperymentalnie, że parametr dotyczący kompromisu między błędem dopasowania a marginesem między klasami będzie miał wartość 5. Dodatkowo zauważono, że im mniejszy parametr regulujący kompromis pomiędzy akceptowanym przez nas błędem dopasowania modelu, a wielkością marginesu pomiędzy klasami, tym większy będzie margines pomiędzy klasami nawet kosztem dużych błędów dla wykrywania wyjątków. W

wyznaczonym klasyfikatorze opartym na modelu wektorów nośnych dla wykrywanych wyjątków bardzo szybko następuje wzrost zarówno błędu klasyfikatora oraz błąd walidacji krzyżowej. Zwiększa się w dużym stopniu liczba wektorów wspierających. Maleje czułość, dokładność i specyficzność algorytmu przy istniejących wyjątkach w analizowanym zbiorze danych. Wyniki dotyczące wyznaczonej dokładności, czułości oraz specyficzności omawianego algorytmu dla zbioru w którym znajdują się wyjątki oraz dla zbioru bez wyjątków pokazano w Tabeli nr 1. Dokładność, czułość oraz specyficzność wyznaczono na podstawie wzorów (14), (15), (16).

$$\frac{FP+FN}{TN+FP+FN+TP} \tag{14}$$

$$\frac{TP}{FN+TP} \tag{15}$$

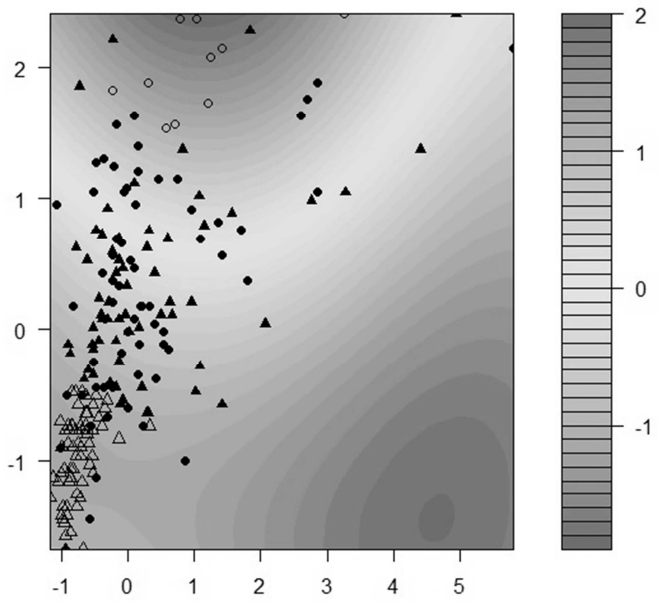
$$\frac{TN}{TN+FP} \tag{16}$$

Tabela. 1. Wyznaczona dokładność, czułość i specyficzność algorytmu SVM dla zbioru A zawierającego wyjątki oraz zbioru B niezawierającego wyjątków.

SVM	Zbiór A	Zbiór B
Dokładność	0.25	0.29
Czułość	0.86	0.71
specyficzność	0.68	0.61

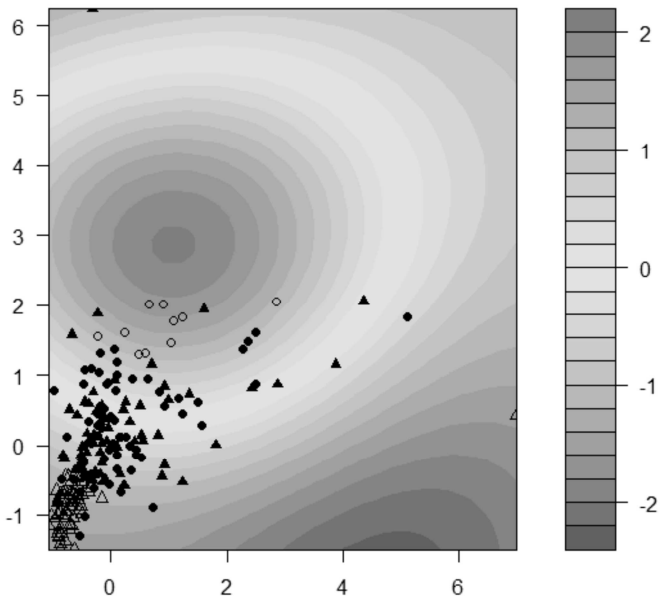
Ilustracje graficzną wyznaczania wyjątków metodą SVM z zaznaczonymi obszarami decyzyjnymi pokazano na Rys. 5.

Na podstawie analizy rys. 5 można stwierdzić, które z obiektów to wyjątki. Łatwo zauważyć, że obiekty podzielono na kółka i trójkąty oraz że mamy dwie klasy decyzyjne zaznaczone kolorami. (odcienie szarości). Kółka, które znajdują się na mocno szarym tle poniżej -1 oraz trójkąty, znajdujące się na mocno szarym tle między 1 a 2 stanowią odchylenia. W bazie istnieją więc poszukiwane wyjątki.



Rys. 4. Ilustracja graficzna obszarów decyzyjnych algorytmu SVM z wyjątkami

Analogiczne wyniki uzyskano dla kolejnego zbioru z repozytorium [17].



Rys. 5. Ilustracja graficzna obszarów decyzyjnych algorytmu SVM z wyraźnym obszarem obiektów będących wyjątkami

Wyznaczone obiekty przy użyciu metody wektorów nośnych jako wyjątki na Rys. 5 i Rys. 6 pokrywają się z wyjątkami wyznaczonymi poprzez inne stosowane klasyfikatory. Należy jednak podkreślić, że graficzna interpretacja podana powyżej z hiperpłaszczyznami jest dokładniejsza.

5 Podsumowanie

Reasumując wyniki badań należy stwierdzić, iż metoda wektorów nośnych może być używana do wykrywania wyjątków w dużych zbiorach danych. SVM lepiej radzi sobie z wykrywaniem w porównaniu do klasyfikatora bayesowskiego czy też k – najbliższych sąsiadów. Istniejące w analizowanym zbiorze danych wyjątki w bardzo dużym stopniu wpływają na błędy klasyfikacji co wykazano w pracy. Dla metody wektorów nośnych uzyskano podczas wykrywania wyjątków najlepsza czułość i dokładność klasyfikatora. W dalsze badania będą skupiały się na stworzeniu funkcji dedykowanych dla klasyfikatora wykrywających wyjątki. Dodatkowo metoda wektorów wspierających będzie użyta w celu wykrycia wyjątków w strumieniach danych.

Bibliografia

[1] Aggarwal, Charu C., Outlier Analysis, Springer, 2013.
[2] Barnett, V., Lewis, T., Outliers in statistical data, Wiley, 1994.
[3] Breuning, M.M., Kriegel, H-P., Ng, R.T., Sander, J., LOF: identifyingdensity-basedlocaloutliers, Proc. ACM SIGMOD Conference on Management of Data, 2000, 93-104.
[4] Chomatek, L. and Duraj, A., Multiobjective genetic algorithm for outliers detection, In: INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on, IEEE, 2017, pp. 379–384.
[5] Duraj, A. and Chomatek, L., Supporting Breast Cancer Diagnosis with Multi-objective Genetic Algorithm for Outlier Detection, In: International Conference on Diagnostics of Processes and Systems, Springer, 2017, pp. 304–315.
[6] Duraj, A. and Krawczyk, A., Finding outliers for large medical datasets, Przeglad Elektrotechniczny, Vol. 86, 2010, pp. 188–191.
[7] Duraj, A. and Szczepaniak, P. S., Information Outliers and Their Detection, In: Information Studies and the Quest for Transdisciplinarity, World Scientific Publishing Company, 2017, pp. 413–437.
[8] Duraj, A., Szczepaniak, P. S., and Ochelska-Mierzejewska, J., Detection of Outlier Information Using Linguistic Summarization, 2016, pp. 101–113.

[9] Duraj, A., Outlier detection in medical data using linguistic summaries, In: INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on, IEEE, 2017, pp. 385–390.
[10] Ester, M., Kriegel, H.-P., Sander, J., Xu, X., et al., A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd, Vol. 96, 1996, pp. 226–231.
[11] Knorr, E.M., Ng, R.T.,Tucakov, V., Distance-basedoutliers: algorithms and applications, VLDB Journal 8, 3-4, 2000, 237-253.
[12] Emets, V. and Rogowski, J., Scattering of acoustical waves by a hard strip and outlier phenomenon, In: INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on, IEEE, 2017, pp. 376–378.
[13] Smolinski, M., Resolving classical concurrency problems using adaptive conflictless scheduling, In: INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on, IEEE, 2017, pp. 397–402.
[14] Smoliński M., Efficient multidisk database storage configuration. In: International Conference: Beyond Databases, Architectures and Structures. Springer, Cham, 2015. pp. 180-189.
[15] Smoliński M., Elimination of task starvation in conflictless scheduling concept. Information Systems in Management, 2016, 5.2, pp. 237-247.
[16] Osowski S, Sieci neuronowe do przetwarzania informacji. Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2006
[17] UC Irvine Machine Learning Repository. http://archive.ics.uci.edu/ml/index.html.
[18] Kumar M.Arun, A hybrid SVM based decision tree, Pattern Recognition, 2010
[19] Vapnik A., The Nature of Statistical Learning Theory (Statistics for Engineering and Information Science), Springer-Verlag, Nowy Jork, 1999

OUTLIERS DETECTION USING SUPPORT VECTOR MACHINE

Summary – Outlier detection in data covers a broad spectrum of science research. In this paper, the author proposes an approach to outlier detection based on support vector machine. In data, an outlier may be considered as a deviation which indicates the existence of outliers. The paper presents the results of tests which were conducted on the set of data from the repository [19].

Keywords: outliers detection, support vector machine