PRACE NAUKOWE AKADEMII EKONOMICZNEJ WE WROCŁAWIU
Nr 1169 2007
TAKSONOMIA 14
Klasyfikacja i analiza danych – teoria i zastosowania

Michał Trzęsiok
Akademia Ekonomiczna w Katowicach

IDENTYFIKACJA OBSERWACJI ODDALONYCH Z WYKORZYSTANIEM METODY WEKTORÓW NOŚNYCH

1. Wstęp

Niezależnie od tego, jak bardzo wyrafinowana metoda zostanie użyta do zbudowania modelu statystycznego, jakość tego modelu zależy wprost od jakości danych wykorzystanych do jego wyznaczenia. Często w rzeczywistych zbiorach danych występują pewne obserwacje, w których wartości opisujących je zmiennych są nietypowe. Wynika to ze specyfiki badanego zjawiska lub też z różnego rodzaju błędów. Ponieważ obserwacje te mogą mieć istotny wpływ na wyniki analizy, wymagają szczególnej uwagi.

Metoda wektorów nośnych (SVM – *support vector machines*), pierwotnie skonstruowana jako metoda dyskryminacji, może zostać przeformułowana tak, aby realizowała także zadanie identyfikowania obserwacji oddalonych. W takim przypadku jej działanie jest podobne w pewnym sensie do metod taksonomicznych (bezwzorcowych), gdyż polega ona na tym, że traktując cały zbiór uczący jak zbiór obiektów należących do jednej klasy, identyfikuje obszar przestrzeni, w którym skupione są poddane analizie obserwacje, tzn. identyfikuje nośnik wielowymiarowego rozkładu, którego realizacjami jest analizowany zbiór danych. Otrzymujemy w ten sposób funkcję przyjmującą wartość zero w obszarach, w których prawdopodobieństwo wystąpienia obserwacji jest bardzo małe, i wartość równą jeden w obszarze będącym nośnikiem rozkładu. Funkcja ta identyfikuje obiekty oddalone.

Dyskryminacyjna metoda wektorów nośnych należy do grupy metod odpornych, tzn. że występowanie obserwacji nietypowych lub błędnie sklasyfikowanych w zbiorze uczącym nie wpływa znacząco na jakość otrzymanego modelu. W artykule przedstawiono metodę SVM przeformułowaną tak, by identyfikowała obserwacje oddalone, oraz próbę empirycznego sprawdzenia, czy przeprowadzenie

wstępnej identyfikacji i usunięcie obserwacji oddalonych poprawia jakość dyskryminacji na zbiorze testowym.

2. Zadanie klasyfikacji bezwzorcowej metodą wektorów nośnych w przypadku jednej klasy

Niech dany będzie zbiór uczący $D=\{\mathbf{x}^1,\ldots,\mathbf{x}^N\}$, gdzie $\mathbf{x}^i \in \mathbf{R}^d$, dla $i=1,\ldots,N$. Zakładamy ponadto, że zbiór uczący to realizacje niezależnych zmiennych losowych o jednakowym rozkładzie P. Traktujemy wszystkie obserwacje jako należące do jednej klasy (klasy obiektów pochodzących z rozkładu P). Zadanie identyfikacji obiektów oddalonych może zostać zrealizowane przez wyestymowanie funkcji gęstości rozkładu P, lecz problem estymowania funkcji gęstości wielowymiarowego rozkładu jest problemem bardzo złożonym. Ponadto nie zawsze problem ten ma rozwiązanie, gdyż zakłada się wtedy, że rozkład ma funkcję gęstości, co nie zawsze jest spełnione (zob. [Smola, Schölkopf 2002]).

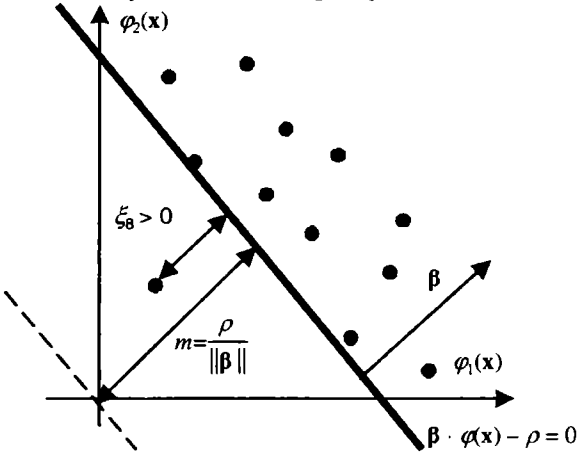
Jak wspomniano we wprowadzeniu – wystarczy znaleźć taki obszar $C \subset \mathbf{R}^d$ wielowymiarowej przestrzeni danych, który spełnia warunek, że niemal wszystkie obserwacje wygenerowane z rozkładu należą do C, a niemal wszystkie obiekty niepochodzące z P należą do dopełnienia zbioru C, czyli do $\mathbf{R}^d \setminus C$. Gdy z góry ustalimy frakcję elementów pochodzących z P, która ma należeć do C, uzasadnione wydaje się nazwanie zbioru C uogólnionym kwantylem rozkładu P. Na tak wyznaczonym zbiorze C definiujemy w sposób naturalny funkcję binarną identyfikującą obserwacje oddalone. Na ogół jednak nawet przy ustalonej frakcji obserwacji zawartych w C, wskazanie takiego zbioru C nie jest jednoznaczne. Wyznaczenie więc zbioru C, korzystając jedynie z informacji zawartej w zbiorze uczącym D, nie jest zadaniem łatwym w pierwotnej przestrzeni danych $\mathbf{R}^d$, choćby ze względu na brak określenia dodatkowych, oczekiwanych własności zbioru takich, jak np. spójność czy wypukłość. Można poszukiwać zbioru C w z góry przyjętej klasie zbiorów pierwotnej przestrzeni $\mathbf{R}^d$, zakładając, że ma on zawierać zbiór uczący D i jednocześnie mieć najmniejszą z możliwych miarę Lebesgue'a (np. w przypadku $\mathbf{R}^3$ poszukiwany byłby zbiór zawierający D, o najmniejszej objętości). Nadal jednak nie można wskazać sposobów decydowania o tym jak, stosownie do problemu, określać przeszukiwaną klasę zbiorów i w jaki sposób znajdować w niej element optymalny.

Ciekawym rozwiązaniem tego problemu jest propozycja wykorzystania techniki stanowiącej podstawę metody wektorów nośnych, która polega na przeniesieniu poszukiwania rozwiązania zadania w przestrzeń o znacznie wyższym wymiarze, przy czym nieliniowa transformacja φ do przestrzeni o większym wymiarze reali-

zowana jest przez wybór pewnej funkcji jądrowej definiującej iloczyn skalarny w nowej przestrzeni cech (zob. [Vapnik 1998]). W tejże nowej przestrzeni wyznaczana jest hiperpłaszczyzna dana równaniem:

$$\boldsymbol{\beta} \cdot \varphi(\mathbf{x}) - \rho = 0, \tag{1}$$

oddzielająca początek układu współrzędnych od punktów będących obrazami obserwacji ze zbioru uczącego. By poszukiwana hiperpłaszczyzna miała jednoznacznie określone położenie w przestrzeni, zakłada się ponadto, że jest ona optymalna w tym sensie, że maksymalizuje szerokość marginesu m (zob. rys. 1). Różnica w stosunku do pierwotnej, dyskryminacyjnej metody wektorów nośnych jest taka, że zamiast oddzielania obiektów różnych klas rozdzielane są punkty zbioru uczącego (traktowane jak elementy jednej klasy) od zbioru jednoelementowego – początku układu współrzędnych. Dodatkowo zakładamy, inaczej niż w dyskryminacji, że hiperpłaszczyzna optymalna nie przechodzi przez środek marginesu, lecz jest położona na jego brzegu, od strony zbioru uczącego (rys. 1).



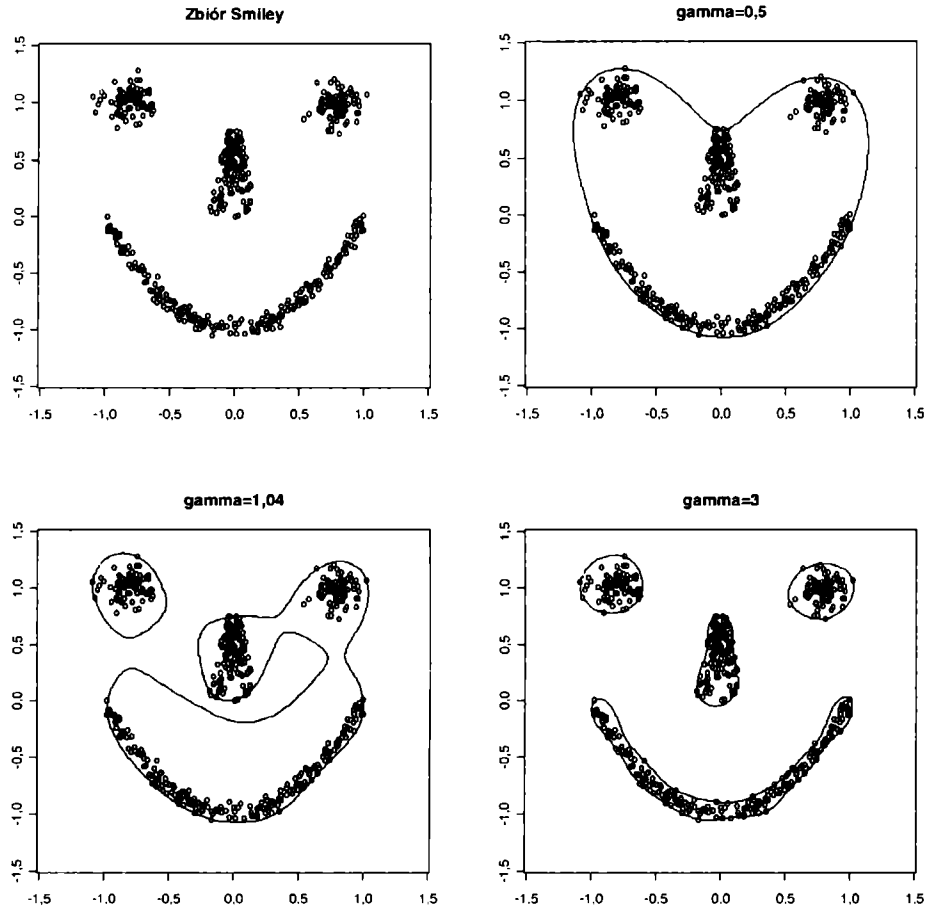
Rys. 1. Optymalna (maksymalizująca margines m) hiperpłaszczyzna oddzielająca początek układu współrzędnych od obrazów obserwacji ze zbioru uczącego w nowej przestrzeni cech. Ponadto na jednym przykładzie zilustrowano konsekwencje wprowadzenia zmiennych ξ_i osłabiających założenie, że wszystkie obserwacje ze zbioru uczącego mają być poprawnie sklasyfikowane przez model
Źródło: opracowanie własne.

Korzyści z zastosowania takiego podejścia są następujące:

1. Zaimplementowane w metodzie SVM zadanie optymalizacyjne rozwiązuje problem znalezienia zbioru o minimalnej „objętości".

2. Wybór klasy zbiorów, w której wyznaczany będzie zbiór C, odbywa się przez wybór funkcji jądrowej, co znacznie ułatwia użytkownikowi posługiwanie się metodą, zmniejsza arbitralność decyzji o wyborze klasy, a jednocześnie gwarantuje, że wybrana klasa pozostaje bardzo liczna i różnorodna.

3. O własnościach takich, jak wypukłość lub spójność, względnie o tym, w jakim stopniu dopuszczamy, by wyznaczony zbiór nie spełniał tych warunków, decyduje użytkownik przez wybór parametrów metody SVM (zob. rys. 2).



Rys. 2. Zaprojektowany do sprawdzania własności metod wielowymiarowej analizy statystycznej zbiór danych *Smiley* (w lewym górnym rogu), wygenerowany komputerowo za pomocą biblioteki `mlbench` pakietu statystycznego R, oraz trzy warianty uogólnionego kwantyla badanego rozkładu, różniące się pod względem spójności. Otrzymano je w wyniku zastosowania metody SVM z parametrem $\nu = 0,01$ oraz różnymi wartościami parametru γ funkcji jądrowej Gaussa
Źródło: opracowanie własne.

4. Mechanizm regularyzacji, kontrolujący kompromis między stopniem dopasowania końcowego modelu do danych ze zbioru uczącego a stopniem złożoności modelu, odpowiedzialny jest za stopień złożoności (gładkości) funkcji opisującej brzeg zbioru C.

Przez analogię do przypadku dyskryminacji zbiorów za pomocą metody wektorów nośnych z parametrem ν przedstawiony problem można formalnie zapisać w postaci zadania optymalizacji wypukłej [Smola, Schölkopf 2002]:

$$\begin{cases} \min\limits_{\boldsymbol{\beta},\beta_0,\boldsymbol{\xi},\rho} \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \rho + \frac{1}{\nu N}\sum\limits_{i=1}^{N}\xi_i, \\ \xi_i \geq 0, \quad \boldsymbol{\beta}\cdot\varphi(\mathbf{x}^{\mathbf{i}}) \geq \rho - \xi_i, \quad i = 1,\ldots,N, \end{cases} \tag{2}$$

gdzie $\nu \in [0,1]$ jest parametrem mechanizmu regularyzacji, za pomocą którego użytkownik określa górną granicę frakcji obiektów ze zbioru uczącego, które mogą zostać sklasyfikowane jako obserwacje oddalone (niepochodzące z rozkładu P). Podobnie jak przy zadaniu dyskryminacji powyższy problem optymalizacyjny można rozwiązać metodą mnożników Lagrange'a, przekształcając go do postaci dualnej, otrzymując końcową postać funkcji identyfikującej obserwacje oddalone:

$$\hat{G}(\mathbf{x}) = \operatorname{sign}[\boldsymbol{\beta}\cdot\varphi(\mathbf{x}) - \rho] = \operatorname{sign}\left[\sum_{i\in I_{SV}} \alpha_i \varphi(\mathbf{x}^{\mathbf{i}})\cdot\varphi(\mathbf{x}) - \rho\right] = \operatorname{sign}\left[\sum_{i\in I_{SV}} \alpha_i \mathrm{K}(\mathbf{x}^{\mathbf{i}},\mathbf{x}) - \rho\right], \tag{3}$$

gdzie przez α_i oznaczono współczynniki Lagrange'a, przez $\mathrm{K}(\mathbf{x}^{\mathbf{i}},\mathbf{x})$ – funkcję jądrową definiującą iloczyn skalarny w nowej przestrzeni cech, a zbiór:

$$I_{SV} = \{i \in \{1,\ldots,N\} : \alpha_i > 0\} = \{i \in \{1,\ldots,N\} : \mathbf{x}^{\mathbf{i}} \text{ jest wektorem nośnym}\},$$

jest zbiorem indeksów wskazujących wektory nośne. Tak zdefiniowana funkcja $\hat{G}(\mathbf{x})$ przyjmuje wartość –1 dla obserwacji oddalonych i wartość 1 dla obserwacji pochodzących z rozkładu P (tj. dokładniej – pochodzących z uogólnionego kwantyla).

Jak wykazano w pracy [Trzęsiok 2004] najlepsze rezultaty w zadaniach dyskryminacji daje metoda wektorów nośnych z funkcją jądrową Gaussa:

$$\mathrm{K}(\mathbf{u},\mathbf{v}) = \exp(-\gamma\|\mathbf{u}-\mathbf{v}\|^2), \tag{4}$$

z odpowiednio dobraną wartością parametru γ oraz wielomianowa funkcja jądrowa:

$$\mathrm{K}(\mathbf{u},\mathbf{v}) = \gamma(\mathbf{u}\cdot\mathbf{v} + \delta)^d, \quad d = 1,2,\ldots, \tag{5}$$

w której kluczowe znaczenie dla jakości modelu ma wybór stopnia wielomianu d.

Rysunek 2. przedstawia zbiór danych *Smiley*, wygenerowany komputerowo za pomocą procedury zawartej w bibliotece `mlbench` pakietu statystycznego R (zob. [Leisch 2004]). Biblioteka ta zawiera wiele różnych zbiorów danych standardowo wykorzystywanych do analizy własności metod wielowymiarowej analizy statystycznej. Zbiór *Smiley* użyto wyłącznie w celu zilustrowania efektu zastosowania metody SVM do wyznaczenia uogólnionego kwantyla dla różnych wartości parametru γ funkcji jądrowej Gaussa.

Na przedstawionym przykładzie widać, że metoda SVM pozwala, przy odpowiednim doborze parametrów, na bardzo precyzyjne wyznaczenie nieliniowej hiperpowierzchni ograniczającej obszar uogólnionego kwantyla, a wybór wartości pa-

rametrów metody bezpośrednio wpływa na spójność obszaru i zdolność modelu do poprawnego identyfikowania nowych obserwacji spoza rozkładu P.

3. Badanie wpływu usunięcia ze zbioru uczącego obserwacji oddalonych na jakość modelu dyskryminacji

W algorytmie dyskryminacyjnej metody wektorów nośnych wprowadzone są zmienne ξ_i uelastyczniające metodę na wypadek występowania w zbiorze uczącym obserwacji błędnie sklasyfikowanych. Z tego względu metoda SVM uznawana jest za metodę odporną na występowanie szumu i obserwacji oddalonych. W szczególności wiadomo, że obserwacje oddalone są przez algorytm metody SVM identyfikowane jako wektory nośne i jako takie mają kluczowe znaczenie dla położenia optymalnej hiperpłaszczyzny rozdzielającej klasy, a tym samym – dla postaci funkcji dyskryminującej. Dodajmy dla porządku, że zbiór wektorów nośnych obejmuje nie tylko obserwacje oddalone, ale również obiekty leżące na brzegu marginesu lub obiekty leżące we wnętrzu marginesu, lecz po „właściwej" stronie hiperpłaszczyzny rozdzielającej (zob. [Cristianini 2000]).

Występowanie w zbiorze uczącym obserwacji oddalonych jest często traktowane przez użytkownika za zjawisko negatywne, mogące niekorzystnie wpływać na zdolność modelu do poprawnego klasyfikowania nowych obiektów. Zasadne więc wydaje się podjęcie próby empirycznego sprawdzenia, czy zidentyfikowanie obserwacji oddalonych metodą wektorów nośnych oraz usunięcie tych obserwacji ze zbioru uczącego spowodują poprawę jakości modelu dyskryminacji.

W analizie wykorzystano cztery zbiory rzeczywistych danych z biblioteki `mlbench`. Są to zbiory: *Sonar* (zawierający 208 obserwacji, reprezentujących 2 klasy, charakteryzowanych przez 61 zmiennych), *Glass* (zawierający 214 obserwacji, reprezentujących 7 klas, charakteryzowanych przez 10 zmiennych), *Satellite* (zawierający 6435 obserwacji, reprezentujących 6 klas, charakteryzowanych przez 37 zmiennych), *Vehicle* (zawierający 846 obserwacji, reprezentujących 4 klasy, charakteryzowanych przez 19 zmiennych).

W każdym ze zbiorów wydzielono na użytek dyskryminacyjnej metody SVM część uczącą (zawierającą 66% obiektów) i testową (pozostałe 34%). Następnie w każdym zbiorze uczącym z osobna dokonano identyfikacji obserwacji oddalonych metodą wektorów nośnych, traktując wszystkie obiekty badanego zbioru jak obserwacje jednej klasy. Na tym etapie parametry metody ustalano przez przeszukanie odpowiednio dużego zakresu ich dopuszczalnych wartości tak, by otrzymać z góry założoną (arbitralnie) frakcję obserwacji oddalonych (uogólniony kwantyl). Zbiór obserwacji oddalonych oznaczono przez O. W kolejnym kroku zbudowano modele dyskryminacji w dwóch wersjach: na pełnym zbiorze uczącym D oraz na zmodyfikowanym zbiorze uczącym $D\backslash O$, powstałym przez usunięcie zidentyfikowanej wcześniej grupy obserwacji oddalonych. Dobór parametrów metody na tym

etapie odbywał się przez przeszukiwanie pewnego zakresu możliwych ich wartości z minimalizacją błędu klasyfikacji liczonego metodą sprawdzania krzyżowego na zbiorze uczącym. W ostatnim etapie badania dokonano oceny jakości modeli w obu wersjach przez obliczenie ich błędów klasyfikacji na wyróżnionym wcześniej zbiorze testowym. Otrzymane wyniki przedstawia tab. 1.

Tabela 1. Błędy klasyfikacji metody wektorów nośnych dla różnych zbiorów danych, dla modeli zbudowanych na całym zbiorze uczącym D (w pierwszym wierszu) i modeli zbudowanych na zbiorze uczącym $D\backslash O$, z którego usunięto zidentyfikowane wcześniej obserwacje oddalone

Nazwa	*Sonar*	*Glass*	*Satellite*	*Vehicle*
Błąd klasyfikacji modelu zbudowanego na D	1,4%	6,6%	11,1%	17,4%
Błąd klasyfikacji modelu zbudowanego na $D\backslash O$	2,8%	5,5%	11,4%	18,7%
Frakcja obserwacji oddalonych	18%	6%	13%	10%

Źródło: opracowanie własne.

Wbrew przypuszczeniu, że obecność obserwacji oddalonych wpływa negatywnie na jakość modelu, okazuje się, iż usunięcie obiektów zidentyfikowanych jako obserwacje oddalone może nawet zwiększyć błąd klasyfikacji. W przeprowadzonym badaniu sytuacja taka wystąpiła aż trzykrotnie. Wprawdzie wzrost błędu był nieznaczny, lecz potwierdza to hipotezę, że dyskryminacyjna metoda wektorów nośnych należy do grupy metod odpornych na występowanie w zbiorze uczącym obserwacji nietypowych i błędnie sklasyfikowanych. Ponadto, ponieważ obserwacje te są w metodzie traktowane jako wektory nośne, czyli obiekty decydujące o postaci funkcji klasyfikującej, w zasadzie można wnioskować, że wręcz nie należy ich usuwać ze zbioru uczącego, gdyż zawierają informacje o specyfice badanego zjawiska. Jedyny przypadek poprawienia jakości modelu w wyniku usunięcia obserwacji oddalonych odnotowano dla zbioru *Glass* i dotyczy sytuacji, gdy usunięte obserwacje nietypowe stanowiły zaledwie 6% liczebności zbioru uczącego. Poprawa jakości modelu nie była jednak znaczna (zaledwie o 1%).

4. Podsumowanie

Metoda wektorów nośnych, oprócz zastosowania w zagadnieniach dyskryminacji i regresji, może być wykorzystana do wyznaczania uogólnionego kwantyla rozkładu i co za tym idzie – do identyfikacji obserwacji oddalonych. Przeprowadzone badania wskazują jednak, że takie stosowanie metody jest zasadne wyłącznie wtedy, gdy identyfikacja obserwacji oddalonych jest celem samym w sobie, tzn. gdy metodami eksploracyjnymi chcemy pozyskać wiedzę o badanym zjawisku z danych, szczególnie uwzględniając informacje zawarte w obserwacjach charakterystycznych, nietypowych. Możliwość uzyskania bardzo różnorodnych postaci obszarów będących uogólnionym kwantylem, nieliniowość funkcji opisujących hiperpowierchnię wyznaczającą ten obszar oraz możliwość otrzymania zbiorów niespójnych to istotne zalety przedstawionego podejścia. Dla badacza jest bardzo

ważne, że przed zastosowaniem metody wektorów nośnych nie musi, a wręcz nie powinien wykorzystywać wstępnych procedur przygotowujących zbiór danych do dalszej analizy, polegających na usuwaniu obiektów nietypowych, gdyż na ogół obiekty te zawierają cenne i unikatowe informacje, które algorytm metody wektorów nośnych potrafi wykorzystać, generując model charakteryzujący się mniejszymi błędami klasyfikacji.

Literatura

Cristianini N., Shawe-Taylor J. (2000), *An Introduction To Support Vector Machines (and other kernel–based learning methods),* Cambridge University Press.

Leisch F., Dimitriadou E. (2004), *The* `mlbench` *Package – a Collection for Artificial and Real–World Machine Learning Benchmarking Problems,* R package, Version 1.0-0, http://cran.R-project.org.

Smola A., Schölkopf B. (2002), *Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond,* MIT Press, Cambridge.

Trzęsiok M. (2004), *Analiza wybranych własności metody dyskryminacji wykorzystującej wektory nośne,* [w:] A.S.·Barczak (red.), *Postępy ekonometrii*, AE, Katowice.

Vapnik V. (1998), *Statistical Learning Theory*, John Wiley & Sons, NY.

NOVELTY DETECTION WITH SUPPORT VECTOR MACHINES

Summary

Support Vector Machines (SVM) are considered as a robust tool for classification. This method is designed for being able to deal with outliers. In the paper Support Vector Machines for single-class problems are presented and the impact of removing previously identified outliers on the classification test set error is analyzed on benchmarking data sets.