Лекция 10. Метод опорных векторов

Метод опорных векторов (SVM, support vector machine) в большинстве случаев позволяет обеспечить хорошее качество распознавания, работая с абстрактной векторной моделью представления данных в метрическом пространстве. Это, как и ранее, означает, что образ описываются вектором признаков $x \in R^n$ и для каждого класса образов задана обучающая выборка $X^{N_j} = \{x^{(j,1)},...,x^{(j,N_j)}\}$, $j = \overline{1,M}$. Задана метрика (функция расстояния), позволяющая сравнивать образы по степени близости.

1. Случай линейно разделимых классов образов

Рассмотрим, первоначально, реализацию метода для случая распознавания двух классов, обладающих свойством **линейной разделимости**. Тогда задача сводится к задаче поиска оптимальной разделяющей гиперплоскости. Очевидно, что вариантов положения такой гиперплоскости в n- мерном пространстве может быть много, поэтому при реализации метода SVM полагают, что граница, локализованная так, чтобы максимизировать зазор (отступ) между классами, способствует более уверенной классификации. Другими словами, ищется разделяющая гиперплоскость, поверхность которой находится на **максимальном** расстоянии от ближайших к ней граничных точек обоих классов из обучающей выборки. Такая гиперплоскость и является в данном случае оптимальной.

Определим обучающие данные в виде смешанной индексированной выборки $X^N = \{x^{(1)},...,x^{(N)}\}$, $D^N = \{d^{(1)},...,d^{(N)}\}$, $d^{(*)} = 1,\ x^{(*)} \in \omega_1, d^{(*)} = -1,\ x^{(*)} \in \omega_2$.

Обозначим каждую пару обучающих данных в виде $\{x^{(i)}, d^{(i)}\}$, $i = \overline{1, N}$. Предположим, что имеется гиперплоскость, однозначно разделяющая данные двух классов (с положительными и отрицательными индексами). Этой гиперплоскости соответствует линейная разделяющая функция, имеющая нулевое значение для всех точек, лежащих на ее поверхности

$$g(x) = x^T w - b_0 = 0, \tag{1}$$

где w – вектор весовых коэффициентов, перпендикулярный к разделяющей гиперплоскости; b_0 – вспомогательная константа.

Тогда условие линейной разделимости образов двух классов можно записать в виде

$$d^{(i)} g(x^{(i)}) = d^{(i)} (x^{(i),T} w - b_0) > 0, \quad i = \overline{1,N}$$

Классификатор на основе метода SVM удобно представить в виде

$$g'(x) = sign(g(x)) = sign(x^T w - b_0), \qquad sign(u) = \begin{cases} 1, & u \geq 0, \\ -1, & u < 0. \end{cases} \tag{2}$$

Очевидно, что изменение масштаба задачи путем умножения вектора w и величины b_0 в (2) на любую константу не меняет результатов классификации. Поэтому удобно выбрать подобную нормировку так, чтобы

$$\min_i d^{(i)} (x^{(i),T} w - b_0) = 1, \; -1 \leq (x^{(i),T} w - b_0) \leq 1. \tag{3}$$

На рис. 1 показан типовой вид разделяющей прямой границы, в которую превращается гиперплоскость в двумерном случае. Расстояние до гиперплоскости определяется длиной перпендикуляра к гиперплоскости, опущенного из начала координат, и равно

$$h = |b_0| / \|w\|.$$

Значениям функции $g'(x) = 1$, соответствует полуплоскость X^+, а значениям $g'(x) = -1$, соответствует полуплоскость X^-.

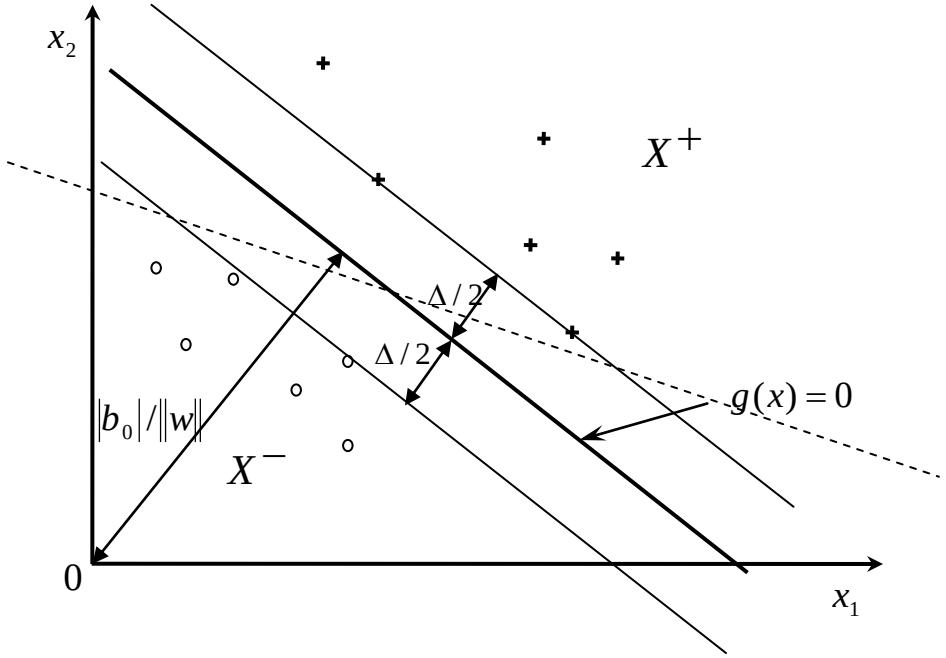


Рис.1. Геометрическая иллюстрация метода SVM

Пусть $x^{(s)}$ и $x^{(t)}$ – две (а, возможно, и более) точки различных классов, являющиеся наиболее близкими к разделяющей гиперплоскости со стороны своего класса. Именно для этих точек в (3) достигается минимум и, соответственно, выполняется

$$g(x^{(s)}) = x^{(s),T} w - b_0 = 1,\ d^{(s)} = 1,\ g(x^{(t)}) = x^{(t),T} w - b_0 = -1,\ d^{(t)} = -1.$$

Геометрически это означает (рис.1), что эти точки лежат на границах «разделяющей полосы», являющихся гиперплоскостями, проходящими по обе стороны искомой разделяющей гиперплоскости параллельно к ней. В случае если обучающая выборка линейно разделима, то гиперплоскости, определяющие границы разделяющей полосы, лучше всего выбирать таким образом, чтобы между ними не лежала ни одна точка обучающей выборки, а расстояние между ними (ширина разделяющей полосы на рис.1) было бы максимальным. Обозначим ширину «разделяющей полосы» величиной Δ.

Таким образом, задача оптимального разделения данных в этом случае сводится к поиску таких векторов, называемых «опорными», а также гиперплоскостей, ближайших к опорным векторам классов, образующих границы разделяющей полосы и, наконец, проходящей параллельно оптимальной разделяющей гиперплоскости.

Очевидно, что величина $\Delta = 2/\|w\|$, т.е. зависит от положения разделяющей гиперплоскости.

Тогда задача сводится к нахождению $\max\limits_{w} \Delta$ для системы неравенств – ограничений

$$d^{(i)} g(x^{(i)}) = d^{(i)}(x^{(i),T} w - b_0) \geq 1, \quad i = \overline{1,N},$$

определяющих принадлежность образов к различным классам и их линейную разделимость. Эта задача может быть сформулирована как задача квадратичного программирования с ограничениями - неравенствами

$$\begin{cases} \dfrac{1}{2} w^T w \to \min, \\ d^{(i)}(x^{(i),T} w - b_0) \geq 1, \quad i = \overline{1,N}, \end{cases} \tag{4}$$

В соответствии с теоремой Куна – Таккера задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа:

$$\begin{cases} L(w, b_0, \lambda) = \dfrac{1}{2} w^T w - \sum\limits_{i=1}^{N} \lambda_i \left[d^{(i)}(x^{(i),T} w - b_0) - 1 \right] \to \min\limits_{w, b_0} \max\limits_{\lambda}, \\ \lambda_i \geq 0, \quad i = \overline{1,N}, \\ \lambda_i = 0 \text{ или } d^{(i)}(x^{(i),T} w - b_0) = 1, \quad i = \overline{1,N}, \end{cases} \tag{5}$$

где $\lambda = (\lambda_1, \ldots \lambda_N)^T$ – вектор переменных, двойственных к ограничению $d^{(i)} g(x^{(i)}) \geq 1$ (множители Лагранжа).

Необходимым условием седловой точки функции Лагранжа является равенство нулю её производных:

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{N} \lambda_i d^{(i)} x^{(i)} = 0, \quad \frac{\partial L}{\partial b_0} = -\sum_{i=1}^{N} \lambda_i d^{(i)} = 0. \tag{6}$$

Из (6) следует, что вектор весовых коэффициентов является линейной комбинацией элементов обучающей выборки, причем тех, для которых $\lambda_i > 0$.

$$w = \sum_{i=1}^{N} \lambda_i d^{(i)} x^{(i)} \,. \tag{7}$$

Именно эти векторы обучающей выборки, для которых $\lambda_i > 0$, принимают участие в формировании решения w и называются **опорными** (support vector)**.**

Для того, чтобы теперь найти величину b_0 достаточно найти w на основе (7) и для произвольный опорного граничного вектора выразить b_0 из равенства

$$b_0 = x^{(i),T} w - d^{(i)}, \quad \lambda_i > 0 \,. \tag{8}$$

На практике для повышения устойчивости рекомендуется брать медиану множества значений b_0, вычисленных по всем граничным опорным векторам

$$b_0 = med\{x^{(i),T} w - d^{(i)},\ \lambda_i > 0\} \,.$$

Таким образом, выражения (7), (8) дают окончательное решение задачи нахождения классификатора в форме (2), оптимального в рамках идеологии SVM

$$g'(x) = sign(\sum_{i=1}^{N} \lambda_i d^{(i)} x^T x^{(i)} - b_0) \,. \tag{9}$$

2. Случай линейной разделимости классов образов с ошибками

На практике случаи линейной разделимости данных гиперплоскостью достаточно редки. Чаще приходится классифицировать образы, которые строго не являются линейно разделимыми. Для обобщения алгоритма SVM на этот случай, вводится допущение появления некоторого количества ошибок на обучающих образах, что позволяет смягчить ограничения в (4).

Задается набор дополнительных переменных $\xi_i \geq 0, i = \overline{1,N}$, характеризующих величину ошибки на классифицируемых объектах выборки $x^{(i)}$, $i = \overline{1,N}$.

Если $\xi_i = 0$, то для $x^{(i)}$ ошибки нет.

Если $0 < \xi_i < 1$, то объект $x^{(i)}$ заступает за границы разделяющей полосы, но корректно классифицируется машиной опорных векторов.

Если же $\xi_i > 1$, то считается, что на объекте $x^{(i)}$ допускается ошибка. С учетом этих допущений задача (5) получит следующий вид:

$$\begin{cases} \dfrac{1}{2} w^T w + C \sum\limits_{i=1}^{n} \xi_i \to \min, \\ d^{(i)} (x^{(i),T} w - b_0) \geq 1 - \xi_i, \quad i = \overline{1,N}, \\ \xi_i \geq 0, \end{cases} \tag{10}$$

где C – «управляющий» параметр метода, который позволяет находить компромисс между максимизацией разделяющей полосы и минимизацией суммарной ошибки.

Функция Лагранжа для задачи (10) имеет вид

$$L(w,b_0,\xi,\lambda,\eta)=\frac{1}{2}w^T w-\sum_{i=1}^{N}\lambda_i\left[d^{(i)}(x^{(i),T}w-b_0)-1+\xi_i\right]-\sum_{i=1}^{N}\xi_i(\eta_i-C)=$$

$$=\frac{1}{2}w^T w-\sum_{i=1}^{N}\lambda_i\left[d^{(i)}(x^{(i),T}w-b_0)-1\right]-\sum_{i=1}^{N}\xi_i(\lambda_i+\eta_i-C),$$

где $\eta=(\eta_1,\dots\eta_N)^T$ – вектор переменных, двойственных к $\xi=(\xi_1,\dots\xi_N)^T$. В данном случае ищется седловая точка для следующей задачи:

$$\begin{cases} L(w,b_0,\xi,\lambda,\eta)\to\min\limits_{w,b_0,\xi}\max\limits_{\lambda,\eta}, \\ \xi_i\ge 0,\ \lambda_i\ge 0,\ \ \eta_i\ge 0, i=\overline{1,N}, \\ \lambda_i=0 \text{ или } \ d^{(i)}(x^{(i),T}w-b_0)=1-\xi_i,\ \ i=\overline{1,N}, \\ \eta_i=0 \text{ или } \ \xi_i=0,\ \ i=\overline{1,N}. \end{cases} \tag{11}$$

Необходимым условием седловой точки функции Лагранжа является равенство нулю её производных:

$$\frac{\partial L}{\partial w}=w-\sum_{i=1}^{N}\lambda_i d^{(i)}x^{(i)}=0,\ \ \frac{\partial L}{\partial b_0}=-\sum_{i=1}^{N}\lambda_i d^{(i)}=0,\ \frac{\partial L}{\partial \xi_i}=-\lambda_i-\eta_i+C=0 \tag{12}$$

Решением задачи, как и ранее, является вектор весовых коэффициентов в виде линейной комбинации элементов обучающей выборки, причем тех, для которых $\lambda_i>0$.

$$w=\sum_{i=1}^{N}\lambda_i d^{(i)}x^{(i)}\ . \tag{13}$$

Именно эти векторы обучающей выборки, для которых $\lambda_i > 0$, принимают участие в формировании решения w и называются **опорными** (support vector)**.**

Для того, чтобы теперь найти величину b_0 аналогично достаточно найти w на основе (13) и для произвольный опорного граничного вектора выразить b_0 из равенства

$$b_0 = x^{(i),T} w - d^{(i)}, \quad \lambda_i > 0 .$$

На практике для повышения устойчивости рекомендуется брать медиану множества значений b_0, вычисленных по всем граничным опорным векторам

$$b_0 = med\{x^{(i),T} w - d^{(i)},\ \lambda_i > 0\} .$$

Анализ также показывает, что $\lambda_i = C - \eta_i \le C,\ i = \overline{1,N}$. При этом допустимы только три возможных комбинации переменных, отвечающих за решение и ограничения:

1. $\lambda_i = 0, \eta_i = C,\ \xi_i = 0,\ d^{(i)} g(x^{(i)}) \ge 1$. Подобное сочетание переменных означает, что вектор $x^{(i)} \in X^N$ классифицирован правильно и не влияет на решение w в (13). Такие векторы называются периферийными.

2. $0 < \lambda_i < C, 0 < \eta_i < C,\ \xi_i = 0,\ d^{(i)} g(x^{(i)}) = 1$. Подобное сочетание переменных означает, что вектор $x^{(i)} \in X^N$ классифицирован правильно, находится на границе разделяющей полосы и определяет решение w в (13). Такие векторы называются опорными граничными.

3. $\lambda_i = C,\ \eta_i = 0,\ \xi_i > 0,\ d^{(i)} g(x^{(i)}) < 1$. Подобное сочетание переменных означает, что вектор $x^{(i)} \in X^N$ либо лежит внутри разделяющей полосы, но классифицируется правильно ($0 < \xi_i < 1\ \ 0 < d^{(i)} g(x^{(i)}) < 1$), либо попадает на границу классов ($\xi_i = 1\ \ d^{(i)} g(x^{(i)}) = 0$) , либо ошибочно отнесен к чужому классу ($\xi_i > 1\ \ d^{(i)} g(x^{(i)}) < 0$) . Такие векторы называются опорными нарушителями.

Таким образом, мы видим, что в полученных выражениях реализуется разделяющей функции реализуется вычисление взвешенной скалярных произведений классифицируемого вектора x и опорных векторов, причем, в данном случае, не только граничных, но и нарушителей

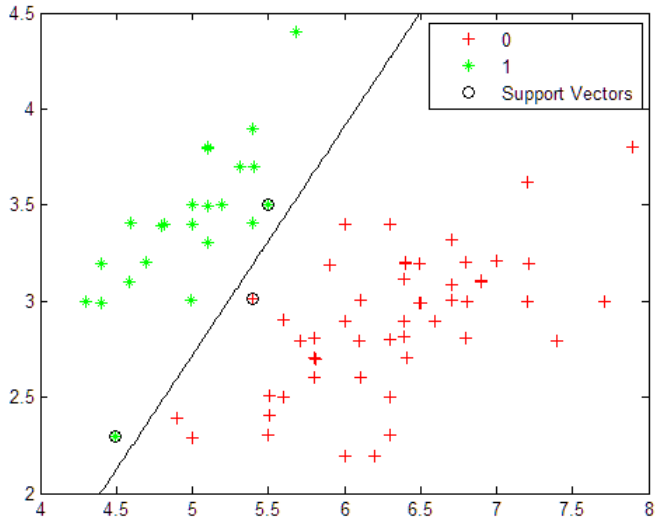
$$g'(x) = sign(\sum_{i=1}^{N} \lambda_i d^{(i)} x^T x^{(i)} - b_0)$$

Последнее обстоятельство может ухудшить работу алгоритма, в ситуациях, когда в обучающей выборке присутствуют «шумовые выбросы».

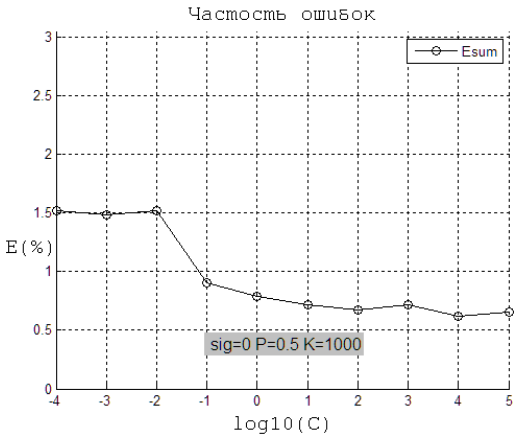
Следует сказать о роли управляющего параметра C. Он играет роль регуляризатора и позволяет находить компромисс между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки. Его введение позволяет повысить устойчивость решения w. выбор конкретного значения может быть проведен на основе метода скользящего контроля.

Если есть основания полагать, что выборка почти линейно разделима, и лишь объекты-выбросы классифицируются неверно, то можно применить фильтрацию выбросов. Сначала задача решается при некотором значении C и из выборки удаляется небольшая доля объектов, имеющих наибольшие значения ошибки ξ_i. После этого задача решается заново по усечённой выборке.

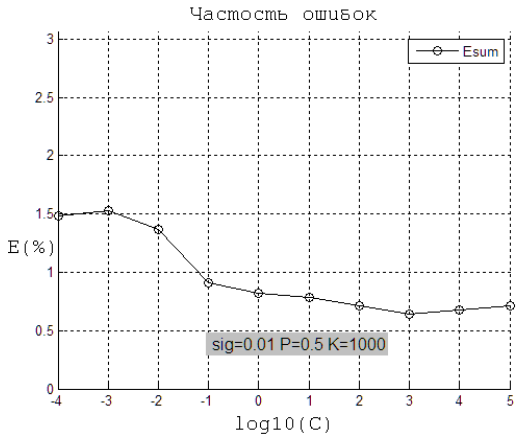
На рис.2 представлены некоторые результаты тестирования алгоритма на данных Ирисы Фишера (fisheriris), включая пример визуализации полученного решения (рис.2а), а также два графика зависимости частости ошибок от значений параметра C в условиях отсутствия (параметр sig=0.0) и наличия искусственного зашумления (параметр sig=0.01) исходных данных (рис.2б,в) при проведении K=1000 экспериментов со случайным разбиением данных на обучающие и тестирующие.



а)



б)



в)

Рис.2. Результаты тестирования алгоритма на основе метода SVM при линейно разделимых классах

3. Случай линейно не разделимых классов образов

Метод SVM может эффективно использоваться и в ситуациях, когда классы образов принципиально линейно не разделимы и разделяющая функция должны быть нелинейной. В подобных ситуациях применяют представленный в предыдущей лекции прием **kernel trick**, в основе которого лежит использование ядра скалярного произведения для перехода в спрямляющее пространство.

Результирующий алгоритм похож на алгоритм линейной классификации, с той лишь разницей, что каждое скалярное произведение заменяется нелинейной функцией ядра (реализующей скалярное произведением в пространстве с большей размерностью). Пожалуй, именно при реализации метода SVM данный прием получил наиболее яркое воплощение. Действительно, полученный ранее классификатор (9) фактически основан на вычислении взвешенной суммы скалярных произведений тестируемого вектора и опорных векторов, а решение оптимизационной задачи можно представить как задачу миминизации квадратичного выпуклого функционала, зависящего от скалярных произведений опорных векторов. Подобная (9) структура классификатора должна сохраниться и после перехода в спрямляющее пространство

$$g'(x) = sign(\sum_{i=1}^{N} \lambda_i d^{(i)} \varphi(x)^T \varphi(x^{(i)}) - b_0).$$

Выполнив замену $K(x, x^{(*)}) = \varphi(x)^T \varphi(x^{(*)})$ и не задавая в явном виде нелинейное преобразование, а ограничиваясь только использованием ядра скалярного произведения, получим алгоритм в следующем виде:

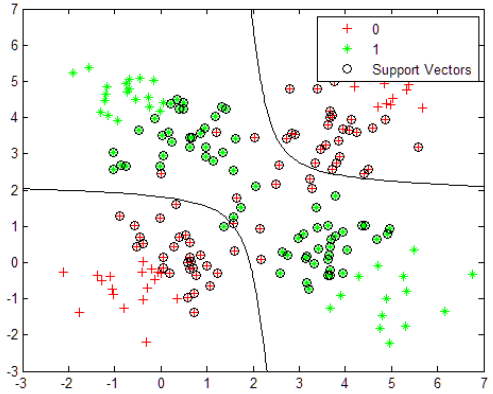
$$g'(x) = sign(\sum_{i=1}^{N} \lambda_i d^{(i)} K(x, x^{(i)}) - b_0). \quad (14)$$

В качестве ядер могут быть использованы любые функции, подходящие по своим свойствам, например, функции, представленные ранее. На рис.3 представлены результаты тестирования алгоритма на основе метода SVM при линейно не разделимых классах на примере «заплетенные восьмерки». Степень «заплетенности» определяется параметром dm, задающим относительный сдвиг центров областей локализации данных в каждом классе (рис.3а,б). Если dm=0, то классы сливаются, если dm становится достаточно большим, то каждый класс состоит из двух не пересекающихся областей эллипсоидальной формы.

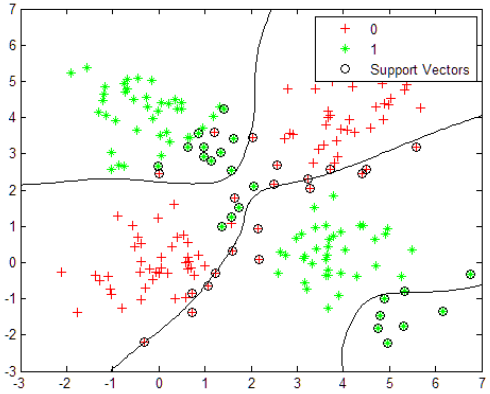
На рис.3 представлены некоторые результаты тестирования программы, включая два примера визуализации полученного решения (рис.3,а,б), полученной при использовании различных функций ядра (а – 'rbf', б – 'mlp')

$$K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right),\quad K(x,z)=\tanh(ax^Tz+b)$$

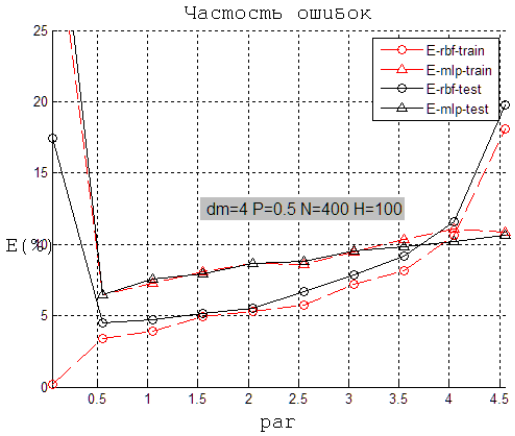
а также совмещенные графики зависимости частости ошибок от (рис.3,в) значений параметра функции ядра (par) и (рис.3,г) относительного объема обучающей выборки (P).



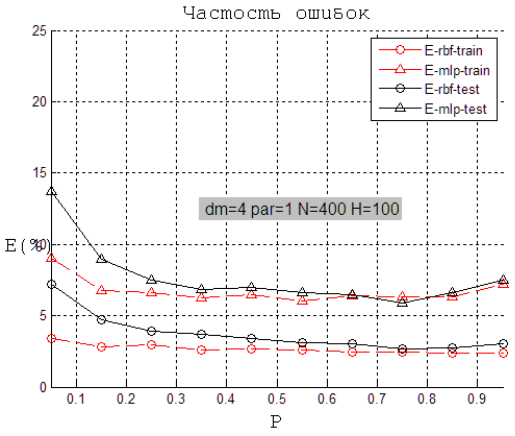
а)



б)



в)



г)

Рис.3. Результаты тестирования алгоритма на основе метода SVM при линейно не разделимых классах

Зависимости на рис.3 представлены для как ошибок, возникающих при тестировании на данных, участвующих в обучении (пунктирные линии), так и для новых данных, не участвующих в процессе обучения.

Здесь мы видим более высокую эффективность алгоритма, использующего функцию ядра в виде радиально-базисной функции. Кроме того, следует отметить достаточно высокую обобщающую способность, показанную алгоритмом (рис.3,г): уже при P>0.3 (объем обучающей выборки больше 100) результаты тестирования на обучающих данных и на новых тестовых данных весьма близки. Это означает, что SVM демонстрирует хорошую устойчивость по отношению к ситуации переобучения.

В случае наличия нескольких классов образов метод опорных векторов применяют путем перехода от задачи классификации на два классов к множественной задаче разбиения на два класса. При этом реализуется два подхода.

Первый предполагает построение бинарных классификаторов, которые различают один класс от остальных (один – против – всех), как это, например, продемонстрировано в первом примере. Количество вычисляемых разделяющих функций в этом случае равно M. Принятие решения осуществляется по принципу максимума значения разделяющей функции для всех возможных сочетаний (победитель получает все). При этом выход функций должен быть прокалиброван для получения сопоставимых оценок.

Второй подход реализует традиционную схему попарного различения классов. Количество вычисляемых разделяющих функций в этом случае равно $M(M-1)/2$. При этом решение принимается в пользу класса, набравшего наибольшее количество голосов при проведении попарного распознавания образа на основе полученной системы разделяющих функций.

Метод SVM в большинстве задач обладает хорошей эффективностью. В тоже время отмечаются его недостатки, в том числе неустойчивость к шумовым искажениям обучающих данных, необходимость подбора параметра C, необходимость обоснования или подбора вида ядра.