УДК 519.688

ОДНОКЛАССОВЫЙ КЛАССИФИКАТОР НА ОСНОВЕ АЛГОРИТМА SVDD

Н. В. Лукьянова

Статья посвящена проблеме классификации сигналов. Представлен классификатор на основе алгоритма SVDD, являющийся модификацией метода опорных векторов. Данный алгоритм позволяет получить гибкую разделяющую гиперсферу вокруг данных тренировочного набора для определения принадлежности тестовых примеров к рассматриваемому классу.

Ключевые слова: *одноклассовый классификатор, описание данных опорными векторами, SVDD, разделяющая гиперсфера*

Введение

В различных отраслях промышленности возникает необходимость решать задачу классификации сигналов, полученных в результате работы диагностических стендов и систем управления технологическими процессами. Существует значительное количество классификаторов, однако далеко не все из них позволяют достигнуть необходимой эффективности при анализе близких по структуре сигналов. Еще одной проблемой, которую необходимо решать при классификации, является принадлежность исследуемого сигнала одновременно к нескольким классам. Использование многоклассового классификатора в данном случае не позволяет полноценно решить поставленную задачу, в то время как одноклассовый классификатор позволяет избежать такого рода проблемы. Это достигается тем, что для каждого исследуемого класса строится свой классифицирующий аппарат, и тестовый сигнал поочередно проверяется на принадлежность к каждому из них.

В данной работе решается задача классификации сигналов, зарегистрированных многоканальными электрокардиографическими приборами. Целью работы является получение классифицирующего программно-аппаратного комплекса, позволяющего эффективно устанавливать имеющийся у обследуемого пациента вид патологии в ходе анализа зарегистрированного многоканального сигнала электрокардиографического прибора.

Постановка задачи

Особенность рассматриваемой задачи заключается в том, что все сигналы зарегистрированы у пациентов, имеющих одну и ту же патологию (все сигналы принадлежат к одной группе), которая известна и обладает ярко выраженными характеристиками. В ходе исследования требуется установить подвид (подгруппу) патологии для каждого сигнала, определение которого в имеющихся диагностических условиях сопряжено со сложностями и в ряде случаев может быть установлено только в ходе оперативного вмешательства.

В настоящей работе рассмотрен одноклассовый классификатор на основе алгоритма *SVDD* (*Support Vector Data Description* – описание данных опорными векторами). Данный метод является модифицированной версией метода опорных векторов [1] и впервые был предложен Д. Таксом и Р. Дуином в 1999 г. [2].

Алгоритм SVDD

Классификатор *SVDD* позволяет выделить набор объектов (опорных векторов) из тренировочного набора данных. Данные размещаются внутри гиперсферы с минимальным объемом ($\vec{\mathbf{a}}$ – центр гиперсферы; R – радиус). При минимизации радиуса гиперсферы снижается вероятность ошибочной идентификации тестируемого объекта с классом, определяемым тренировочным набором. Для того чтобы разрешить наличие исключений в тренировочном наборе, вводится набор параметров ξ_i и некоторая константа C, при помощи которых можно найти компромисс между объемом сферы и полученной ошибкой. Константа C показывает количество векторов, которые могут выйти за границу гиперсферы. Используя ограничение, при котором почти все объекты будут лежать внутри сферы:

$$\left\|\vec{\mathbf{x}}_i - \vec{\mathbf{a}}\right\|^2 \le R^2 + \xi_i;\ \xi_i \ge 0\ \forall i,$$

где $\vec{\mathbf{x}}_i$ – i-й пример из обучающей выборки, получим функцию ошибки, которую необходимо минимизировать:

$$F(R, \vec{\mathbf{a}}) = R^2 + C\sum_i \xi_i .$$

Данное выражение может быть преобразовано путем введения множителей Лагранжа α_i, γ_i и построения лагранжиана:

$$L(R,\vec{\mathbf{a}},\xi,\alpha,\gamma) = R^2 + C\sum_i \xi_i - \\ -\sum_i \alpha_i (R^2 + \xi_i - (\left\|\vec{\mathbf{x}}_i\right\|^2 - 2\vec{\mathbf{a}}\vec{\mathbf{x}}_i + \left\|\vec{\mathbf{a}}\right\|^2)) - \sum_i \gamma_i \xi_i. \quad (1)$$

Лагранжиан L необходимо минимизировать по отношению к величинам R, $\vec{\mathbf{a}}$, ξ и максимизировать по отношению к α и γ. Приравняв частные производные выражения (1) по R, $\vec{\mathbf{a}}$ и ξ к нулю, получим следующие ограничения:

$$\frac{\partial L}{\partial R} = 0: \quad \sum_i \alpha_i = 1;$$

$$\frac{\partial L}{\partial \vec{\mathbf{a}}} = 0: \quad \vec{\mathbf{a}} = \sum_i \alpha_i \vec{x}_i;$$

$$\frac{\partial L}{\partial \xi_{\mathrm{i}}} = 0: \quad C - \alpha_i - \gamma_i = 0.$$

Из последнего уравнения следует, что $\alpha_i = C - \gamma_i$, а так как $\alpha_i \ge 0, \gamma_i \ge 0$, то множитель Лагранжа γ_i может быть исключен в случае следующего ограничения:

$$0 \le \alpha_i \le C .$$

Применяя все полученные ограничения к выражению (1), лагранжиан L примет следующий вид:

$$L = \sum_i \alpha_i (\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_i) - \sum_{i,j} \alpha_i \alpha_j (\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j). \quad (2)$$

Максимизация выражения (2) позволит определить набор α_i [3]. В случае, если $\vec{\mathbf{x}}_i$ удовлетворяет неравенству $\left\|\vec{\mathbf{x}}_i - \vec{\mathbf{a}}\right\|^2 < R^2 + \xi_i$, то множитель Лагранжа $\alpha_i = 0$. Для объектов, удовлетворяющих выражению $\left\|\vec{\mathbf{x}}_i - \vec{\mathbf{a}}\right\|^2 = R^2 + \xi_i$, множитель Лагранжа $\alpha_i > 0$. Согласно алгоритму *SVDD* для описания класса необходимо выбрать только опорные векторы (*support vectors*), для которых $\alpha_i > 0$.

Получив набор α_i, можно определить неравенство, задающее границу для задачи одноклассовой классификации. Тестовый вектор $\vec{\mathbf{z}}$ принадлежит к исследуемому классу, если выполняется неравенство

$$\left\|\vec{\mathbf{z}} - \vec{\mathbf{a}}\right\|^2 = (\vec{\mathbf{z}} \cdot \vec{\mathbf{z}}) - 2\sum_i \alpha_i (\vec{\mathbf{z}} \cdot \vec{\mathbf{x}}_i) + \sum_{i,j} \alpha_i \alpha_j (\vec{\mathbf{x}}_i \cdot \vec{\mathbf{x}}_j) \le R^2. \quad (3)$$

Гиперсфера – очень жесткая граница вокруг данных и обычно не дает правильного представления об их структуре. Идея метода *SVDD* – нелинейное отображение данных из тренировочного набора в пространство с большей размерностью и построение разделяющей гиперплоскости в этом пространстве. Таким образом, можно получить нелинейные границы в исходном пространстве. С использованием функции ядра можно вычислить разделяющую гиперплоскость без конкретного отображения данных в пространство большей размерности.

Функцией ядра может быть любая функция, удовлетворяющая теореме Мерсера. В данной работе использовалось ядро Гаусса [4]:

$$K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \exp\left(-\frac{\left|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j\right|^2}{s^2}\right),$$

где s – параметр, подбираемый экспериментально.

Заменив скалярные произведения в выражении (2), получим новое представление для лагранжиана:

$$L = -\sum_{i,j} \alpha_i \alpha_j K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j).$$

Неравенство, задающее границу для задачи одноклассовой классификации, также примет другой вид:

$$\sum_i \alpha_i \exp\left(-\frac{\left|\vec{\mathbf{z}} - \vec{\mathbf{x}}_i\right|^2}{s^2}\right) > \frac{1}{2}\left(1 + \sum_{i,j} \alpha_i, \alpha_j K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) - R^2\right),$$

из которого видно, что граница строго зависит от параметра *s*. При малых значениях *s* обычно все объекты представляют собой тренировочный набор. Для больших значений *s* получаем границу в виде гиперсферы.

С увеличением значения параметра *s* число векторов с ненулевыми значениями α_i уменьшается. Когда тренировочный набор представляет собой нормальное распределение данных, количество тренировочных векторов позволяет дать оценку ошибки при проверке очередного вектора. С увеличением *s* растет число векторов, и, соответственно, уменьшается ошибка. Однако, чем шире граница допуска новых объектов, тем больше риск принять ложный объект. Параметры *s* и *C* подбираются экспериментально для каждой из обучающих и тестовых выборок.

Проведение исследования для определения эффективности классификатора

Все используемые в работе сигналы были зарегистрированы при обследовании пациентов прибором *Cardiag-112.2* в отделении неинвазивной аритмологии Научного центра сердечно-сосудистой хирургии (НЦССХ) им. А.Н. Бакулева РАМН.

Исследование проводилось для двух групп сигналов. Первая группа сигналов представляла собой простую электрофизиологическую модель, вторая – сложную. Для уточнения вида имеющейся сердечно-сосудистой патологии данные обеих групп пациентов необходимо было разделить на две подгруппы (свои для каждой группы).

Перед классификацией была сокращена размерность сигналов. Многоканальный сигнал сокращенной размерности представляет собой набор главных компонент, полученных для одноканальных составляющих сигнала

$$\vec{\mathbf{y}} = \left[y_{1,1},...,y_{1,l},y_{2,1},...,y_{2,l},....,y_{80,1},...,y_{80,l} \right],$$

где l – количество главных компонент, взятых для каждого одноканального сигнала; $y_{i,j}$ – j-я главная компонента i-го сигнала. В случае, если число l велико и длина нового сигнала значительна, можно провести повторное преобразование, вычислив главные компоненты сформированного сигнала.

Для проверки работы данного классификатора были сформированы обучающая, тестовая и контрольная выборки для каждой группы пациентов. В ходе исследования параметр *s* менялся от 0,2 до 1,4. Параметр *C* изменялся в диапазоне от 1 до 5.

Результаты классификации многоканальных сигналов

Для первой группы пациентов в ходе исследования были получены наиболее оптимальные результаты при *C*=2 и *s*=0,8 для 1 класса и *C*=2 и *s*=1,2 для 2 класса (по данным тестовой выборки, табл. 1).

Для второй группы пациентов в ходе исследования были получены наиболее оптимальные результаты при *C*=1 и *s*=0,6 для 1 класса и *C*=1 и *s*=0,8 для 2 класса (табл. 2).

Результаты, полученные для сигналов второй группы, заметно хуже, чем для первой. Это обусловлено тем, что во второй группе

Таблица 1

Результаты классификации сигналов первой группы

Характеристики	Группа 1	
	Класс 1	Класс 2
	C = 2, *s* = 0,8	*C* = 2, *s* = 1,2
Обучающая выборка (количество сигналов)	10	10
Контрольная выборка (количество сигналов)	10	10
Чувствительность классификатора, %	90	90
Специфичность классификатора, %	90	80

Таблица 2

Результаты классификации сигналов второй группы

Характеристики	Группа 2	
	Класс 1	Класс 2
	C = 1, *s* = 0,6	*C* = 1, *s* = 0,8
Обучающая выборка (количество сигналов)	26	23
Контрольная выборка (количество сигналов)	16	12
Чувствительность классификатора, %	75	70
Специфичность классификатора, %	72	68

были представлены сигналы сложной структуры, зарегистрированные у пациентов с большим количеством сопутствующих заболеваний, что сильно искажает сигнал. Корректная классификация таких сигналов крайне сложна, и в ряде случаев окончательный подвид патологии можно установить только в ходе оперативного вмешательства.

Таким образом, разработанный классификатор позволяет значительно повысить эффективность диагностирования по имеющимся данным многоканальной ЭКГ и уменьшить риск для пациента, связанный с проведением дополнительных клинических обследований.

Заключение

В работе был рассмотрен статистический классификатор на основе алгоритма *SVDD* и проведено исследование эффективности его применения при анализе многоканальных сигналов электрокардиографических приборов. В ходе исследования экспериментальным образом был получен ряд необходимых для классификации параметров.

Рассмотренный классификатор эффективен при анализе сигналов как простой (до 90 % правильной классификации), так и сложной структуры (около 70 %) уже на небольших обучающих выборках. Такая эффективность при использовании небольшого количества примеров в сформированных выборках позволяет использовать данный классификатор для уточнения класса объекта уже в начале проводимых исследований.

Апробация классификатора на основе алгоритма SVDD проводилась в отделении неинвазивной аритмологии НЦССХ им. А.Н. Бакулева. Вид патологии, установленный при анализе всех рассмотренных в работе сигналов, впоследствии был подтвержден в ходе внутрисердечных исследований.

Список литературы

1. Vapnik V.N. Statistical learning Theory. – Wiley, 1998. – 740 p.
2. Tax D., Duin R. Support vector data descrip-tion // Pattern Recognition Letters. 1999. Vol. 20. P. 1191–1199.
3. Акулич И.Л. Математическое программирование в примерах и задачах. – М.: Высшая школа, 1986. – 318 с.
4. Vilaplama V., Marques F. Support vector data description based on PCA features for face detection // 13th European Signal Processing Conference, September, 2005. P. 115 – 119.

Материал поступил в редакцию 5.12.2010 г.

ЛУКЬЯНОВА Наталия Владимировна

E-mail: **luck-va@msiu.ru**
Тел. **+7 (495) 620-39-39**

Старший преподаватель кафедры информационных систем и технологий МГИУ. Область научных интересов – обработка и анализ сигналов, нейронные сети. Автор 10 научных трудов.