УДК 519.7

ОБ ОДНОМ ПОДХОДЕ К СИНТЕЗУ «ПРОСТЫХ» АЛГОРИТМОВ КЛАССИФИКАЦИИ

Тихомиров М. Г.

Московский Педагогический Государственный Университет,
Математический Факультет
ул. Малая Пироговская, 1г. Москва, Россия, 119882
e-mail: *MTIKHOMIROV@ELCO.RU*

Abstract

The analysis of the consequent achievements and publications devoted to the issue of synthesis of high-performance and exact algorithms of classification reveals that the issue of the problem-solving procedure when a «simple» decision rule is required still remains open. The method of attack, based on the synthesis of the «virtual» precedents for classification, of which high-accuracy but not inadmissibly «complicated» algorithms are used, is offered in this article. The main concepts of the approach are demonstrated with the mode! examples.

Введение

Для многих важных прикладных задач классификации в качестве существенного дополнительного требования выступает необходимость построения в некотором смысле «простого» решения. Например, оценки надежности банков должны вычисляться по явно выписанным формулам, в которых в качестве переменных выступают балансовые показатели. В строительных нормах и правилах (СНиПах) также используются лишь простейшие комбинации параметров, сравниваемые с нормативными порогами. Аналогичным образом интерпретируются результаты анализов крови человека и т.д.

При рассмотрении задачи классификации как задачи обучения по прецедентам достаточно естественным представляется подход, при котором с самого начала фиксируется подходящее по сложности семейство алгоритмов и в его рамках ищется оптимальное решение. Использование этого подхода на самом деле основывается на предположении о том, что в обучающей совокупности классы представлены в достаточной степени равноправно как в количественном смысле, так и в смысле геометрической равномерности заполнения пространства допустимых объектов. На рис. 1 представлен пример плоской задачи, на котором данный подход дает содержательно верное решение.

Возможны ситуации, когда классы представлены существенно неразномерно и при этом объекты обучающей совокупности неравномерно распределены по компакту допустимых объектов. Если в обучающем множестве на границе компакта увеличить количество объектов более полно представленного класса, то описанное ранее «простое» решение будет давать много ошибок даже на обучении. На рис. 2 представлена подобная ситуация.

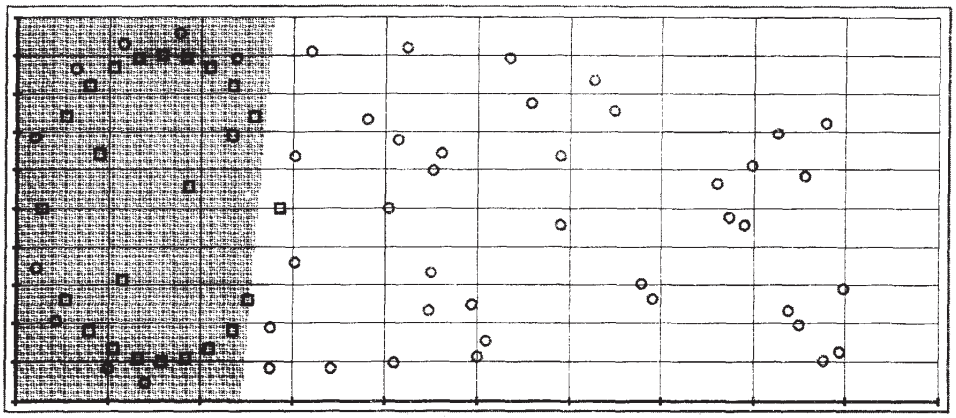

Рис. 1. «Простое» решение с фиксированным по сложности семейством алгоритмов

При дальнейшем увеличении количества прецедентов более полно представленного класса «простое» решение перестает «замечать» мало представленный класс, относя все прецеденты к одному классу. В рассматриваемом примере добавление даже двух объектов приводит к данной ситуации (рис. 3).

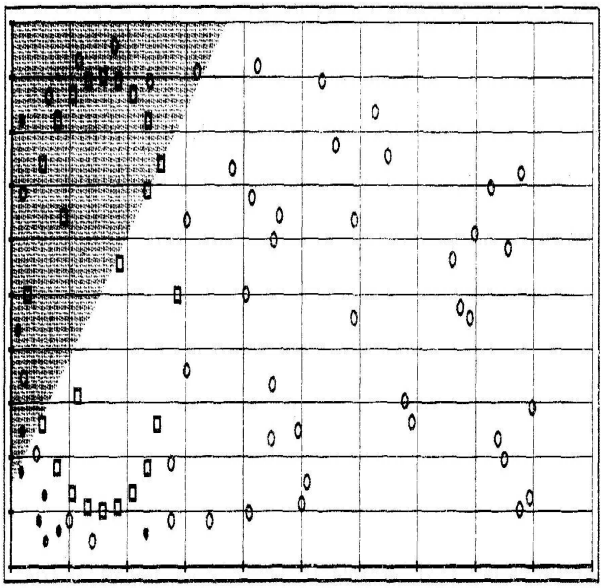

Рис. 2. Влияние увеличения количества объектов более полно представленного класса на границе компакта

Целью настоящей работы является описание на принципиальном уровне альтернативного подхода, пригодного для решения задач классификации в ситуациях, когда требуется построить в некотором смысле «простое» решающее правило.

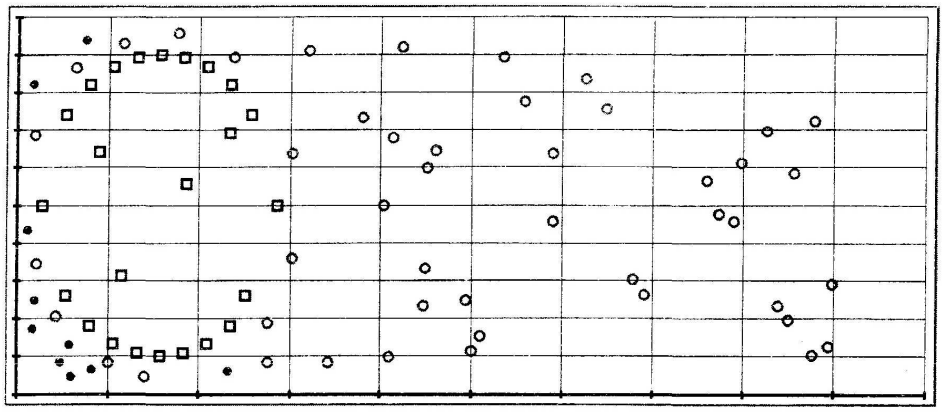

Рис. 3. Пример «простого» решения, относящего все объекты к одному классу

Синтез «простых» алгоритмов классификации

В качестве исходной гипотезы предлагаемого подхода принимается положение о том, что при снятии ограничения на сложность может быть построен, например - с помощью конструкций алгебраического подхода, алгоритм классификации достаточно высокого качества. Суть предлагаемого подхода сводится к тому, что оптимизация в рамках «простого» семейства алгоритмов должна проводиться не на исходном множестве обучающих объектов, но на расширенном множестве объектов, включающем в себя «виртуальные» прецеденты, принадлежность которых к классам определяется «сложным» алгоритмом.

Решение задачи предлагаемым способом состоит из четырех этапов.

1 этап. Получение точного «сложного» решения.

2 этап. Синтез дополнительных «виртуальных» прецедентов «равномерно» распределенных в пространстве объектов.

3 этап. Разнесение полученных на 1 этапе «сложным» решением виртуальных объектов по классам.

4 этап. Синтез оптимального «простого» решения с учетом добавленных «виртуальных» прецедентов.

Точное «сложное» решение может быть получено множеством различных способов, в том числе, с помощью алгебраического подхода [1, 2, 3. 4, 5]. В качестве примера ниже рассматривается случай плоской задачи. В модельных ситуациях как недопустимое по сложности решающее правило используется квадратичное, как «простое» - линейное. Для нахождения границ классов на плоскости используется SVM (support vector machines) [6, 7]. На рис. 4 представлено точное «сложное» решение. При использовании SVM применяется переход в спрямляющее пятимерное пространство,

координаты объектов в котором x, y, x^2, y^2, xy, где x и y соответствующие координаты точек в двумерном пространстве.

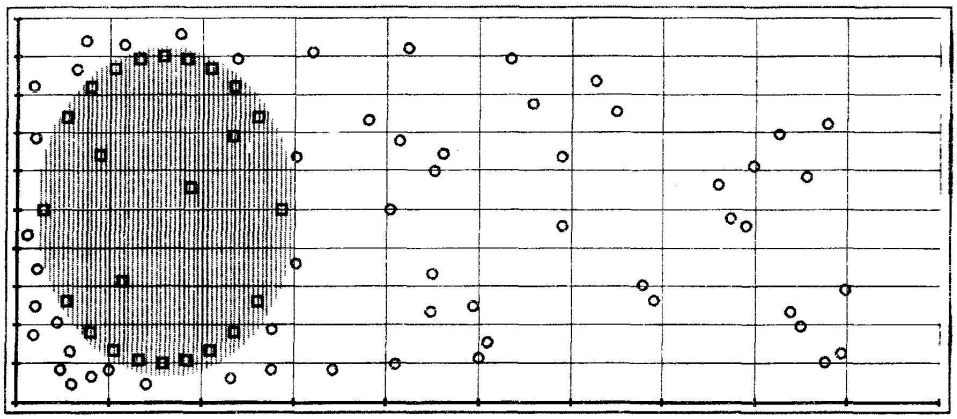

Рис. 4. Точное «сложное» решение

Зачастую, при использовании подхода, при котором сразу фиксируется подходящее по сложности семейство алгоритмов, получается решение, дающее малое количество ошибок на обучающем множестве. В описываемой ситуации, когда классы представлены существенно неравномерно в количественном смысле, это решение относит все точки к более представленному классу (рис. 3). Такое решение, хотя и дает малое количество ошибок на обучающем множестве, получается содержательно неверным. Это решение просто игнорирует слабо представленный класс.

Используя предлагаемый подход, на расширенном множестве объектов мы находим «простое» решение, позволяющее классифицировать новые объекты с помощью удовлетворительного по сложности алгоритма.

Рассмотрим более детально каждый из этапов решения задачи.

I этап.

На первом этапе находим точное «сложное» решение, представленное на рис. 3. Полученный алгоритм будет использоваться на третьем этапе для определения принадлежности «виртуальных» объектов к классам.

II этап.

Для синтеза дополнительных «виртуальных» прецедентов также можно использовать различные подходы, одним из которых является увеличение весов объектов мало представленного класса. Этот подход имеет один существенный недостаток. В случае неравномерного распределения объектов малого класса эта неравномерность сохраняется.

Для получения более точного решения предполагается добавлять «виртуальные» объекты с учетом требования равномерного заполнения всего рассматриваемого компакта как исходными, так и «виртуальными» прецедентами.

Для выбора и оценки алгоритма, добавляющего дополнительные «виртуальные» прецеденты, использовались следующие алгоритмы:

Первый вариант.

а) компакт разбивался сеткой с малым шагом;
б) для каждого объекта находился ближайший узел сетки и расстояние между ними;
в) из этих расстояний выбиралось большее и в качестве «виртуальных» прецедентов добавлялся соответствующий узел сетки.

Второй вариант.

а) компакт разбивался сеткой с малым шагом;
б) поочередно добавлялся каждый из узлов сетки;
в) на новой конфигурации объектов запускался 1 вариант алгоритма и находилось наибольшее расстояние до узлов сетки;
г) из этих расстояний выбиралось минимальное и в качестве «виртуальных» прецедентов добавлялся узел сетки, соответствующий этому расстоянию.

Второй вариант давал более точные результаты только на малом количестве прецедентов (менее 10), при большем количестве объектов результаты двух способов совпадали. Поэтому для оценки описанных ниже алгоритмов добавления точек использовался первый вариант, превосходящий второй по скорости. Два данных способа также можно использовать для добавления «виртуальных» прецедентов, но даже в рассматриваемом простом случае плоской задачи эти алгоритмы требуют больших временных затрат.

Поэтому, при добавлении новых «виртуальных» объектов предлагается использовать следующий алгоритм. Рассматриваемый компакт разбивается на треугольники, т.е. строится триангуляция Делоне. Выбирается треугольник с наибольшей площадью. Одним из описанных ниже способом находится «центр» треугольника. Найденная точка добавляется в множество начальных и «виртуальных» объектов. Цикл повторяется до тех пор, пока не будет достигнута необходимая равномерность распределения прецедентов. Критерием качества в данном случае является наименьшее значение расстояния между прецедентами и узлами сетки.

При поиске «центров» треугольников используются следующие алгоритмы:

1) «центром» треугольника является центр вписанной окружности треугольника;
2) «центром» треугольника является центр тяжести треугольника;
3) «центром» треугольника является точка, координаты которой определяются описанным ниже алгоритмом T.

Алгоритм T

Дан треугольник ABC. Целью работы алгоритма является добавление точки X, так, чтобы круги с вершинами в A,B,C и X радиуса r полностью покрыли треугольник ABC, при этом r должно быть минимальным (рис. 5).

Для решения данной задачи необходимо найти окружность, описанную вокруг незакрытой тремя окружностями, расположенными в вершинах, части треугольника. При нахождении этой окружности мы руководствовались утверждением, что данная окружность будет описанной вокруг одного из треугольников, образуемых поочередным соединением вершин непокрытой части треугольника.

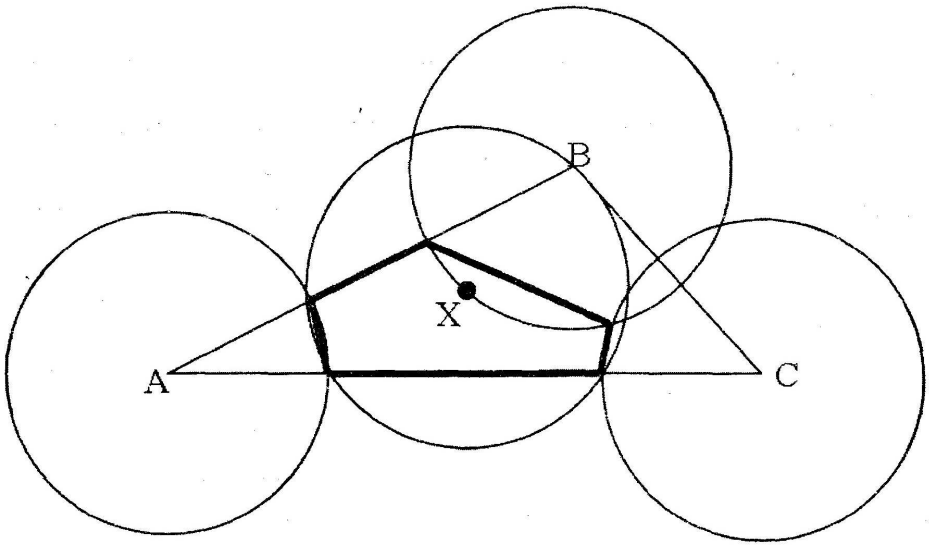

Рис. 5. Нахождение «центра» треугольника

При нахождении четвертой окружности максимальное количество рассматриваемых треугольников равно двадцати, так как максимальное количество вершин непокрытой части треугольника равно шести. Также для треугольных треугольников возможна ситуация, при которой у треугольника остаются две непокрытых области, в данном случае невозможно покрыть треугольник четырьмя окружностями равного радиуса. В данной ситуации мы брали большую из непокрытых частей.

Алгоритм нахождения «центра» треугольника

r_1 и r_2 - значения вспомогательных радиусов;
$r^{'}$- радиус окружностей, расположенных в вершинах треугольника;
$\rho(r^{'})$ - радиус окружности, описанной вокруг непокрытой части треугольника;
Условием окончания работы алгоритма является:

$$|\rho(r^{'}) - r^{'}| < 0{,}001 \cdot \min\left[\rho(r^{'}), r^{'}\right]$$

1) устанавливаем начальное значение радиуса r_1 равным 0,
2) выбираем наибольшую сторону треугольника и устанавливаем начальное значение радиуса r_2 равным длине этой стороны.
3) устанавливаем значение $r^{'}$ равным полусумме r_1 и r_2,
4) находим значение $\rho(r^{'})$.
5) в случае если $\rho(r^{'}) > r^{'}$ устанавливаем $r_1 = r^{'}$, если $\rho(r^{'}) < r^{'}$ устанавливаем $r_2 = r^{'}$,
6) проверяем условие окончания цикла и в случае его не соблюдения повторяем цикл с третьего пункта.

Результаты добавления дополнительных «виртуальных» прецедентов с помощью этого алгоритма представлены на рис. 6.

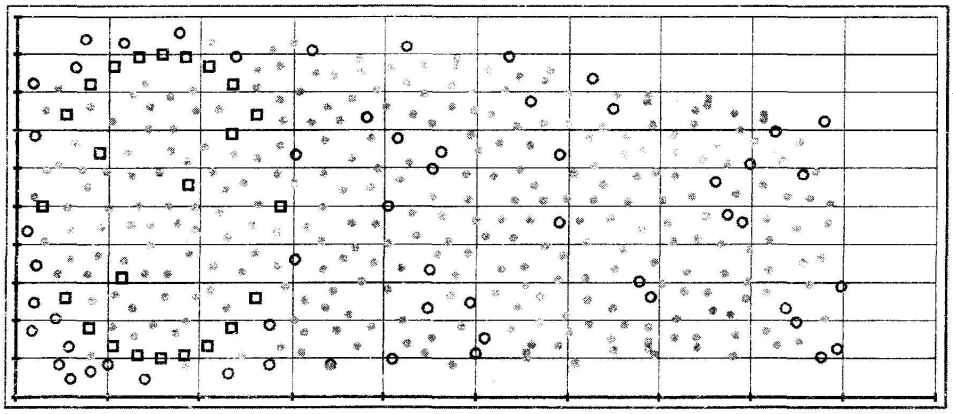

Рис. 6. Результат добавления «виртуальных» прецедентов

III этап.

На этом этапе определяется принадлежность к классам добавленных ранее «виртуальных» прецедентов. Для этого используется полученное на первом этапе «сложное» решение. Распределенное по классам множество начальных и «виртуальных» прецедентов представлено на рис. 7.

IV этап.

На последнем четвертом этапе с помощью SVM синтезируется оптимальное «простое» решение, с одной стороны отвечающее ограничениям по сложности, с другой стороны дающее возможность классифицировать новые объекты. Полученное в итоге «простое» решение представлено на рис. 8.

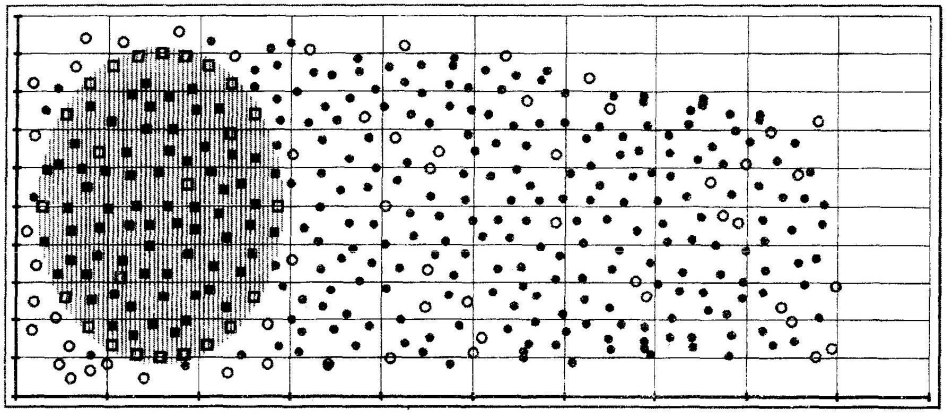

Рис. 7. Распределение «виртуальных» прецедентов по классам

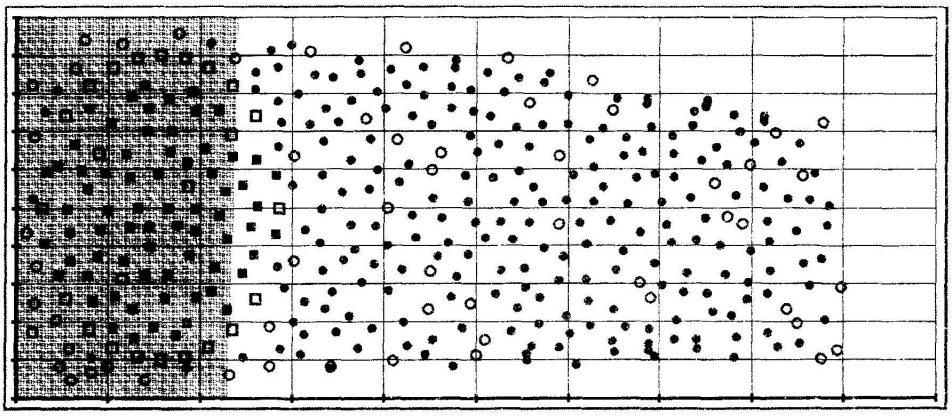

Рис. 8. Итоговое «простое» решение

Заключение

*Основным результатом данной статьи является*новый метод построения «простых» алгоритмов классификации на основе синтеза «виртуальных» прецедентов и результатов их классификации с помощью «сложных» алгоритмов. Работоспособность предлагаемого метода продемонстрирована на примере построения линейных дискриминантных поверхностей. Основным направлением дальнейших исследований будет расширение семейства алгоритмов классификации, допускающих применение разрабатываемого подхода.

В заключение, автор статьи выражает благодарность своему научному руководителю Рудакову К.В.

Список литературы

1. Журавлев Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации Проблемы кибернетики. Вып. 33. М.: Наука, 1978. С. 5 - 68.
2. Рудаков К.В. Об алгебраической теории универсальных и локальных ограничений для задач классификации Распознавание, классификация, прогноз. М.: Наука, 1989. С. 176 - 201.
3. Рудаков К.В., Чехович Ю.В. Критерии полноты моделей алгоритмов и семейств решающих правил для задач классификации с теоретико-множественными ограничениями Доклады Академии наук, 2004 г. том 394, № 4. С. 459 - 461.
4. Рудаков К.В., Чехович Ю.В. Критерии полноты семейств корректирующих операций и моделей алгоритмических операторов для задач классификации с теоретико-множественными ограничениями Доклады Академии наук, 2004 г. том 395, № 6. С. 749 - 750.
5. Рудаков К,В., Чехович Ю.В. Критерии полноты для задач классификации с теоретико-множественными ограничениями Журнал вычислительной математики и математической физики. 2005. том 45. № 2. С. 344 - 353.
6. B.E.Boser, I.M.Guyon, V.N.Vapnik, A training algorithm for optimal margin classifiers, in Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, ACM Press, 1992.
7. C.Cortes, V.Vapnik, Support Vector Networks, Machine Learning 20(3), 1995.