

УДК 004.93'1

В.И. Ширяев, Б.М. Кувшинов

Южно-Уральский государственный университет, г. Челябинск, Россия

Использование адаптивных методов распознавания образов в задачах принятия решений*

Предложен подход к решению задачи автоматической классификации многопараметрических объектов методами распознавания образов, учитывающий неполноту знаний экспертов на этапе построения системы распознавания. Он позволяет совместить использование экспертных суждений и возможности математических методов анализа данных в рамках единого процесса обучения системы распознавания. В основу системы классификации положены комитетные решающие правила, для которых предложен алгоритм обучения, учитывающий неопределенность априорной информации. Эффективность разработанных алгоритмов проверена при решении практической медицинской задачи оценивания тяжести гипертрофической кардиомиопатии.

Введение

Системы автоматической классификации, основанные на методах распознавания образов, получили широкое применение в самых разнообразных системах управления [1]. Они представляют собой эффективный инструмент автоматизации процесса принятия решений. Такие системы позволяют формализовать знания экспертов, принимающих решения об управлении какими-либо объектами, и автоматизировать этот процесс. Здесь под решением эксперта понимается отнесение объекта, обладающего определенным набором значений параметров, к одному из нескольких классов. При этом параметры объекта могут быть измерены или оценены и характеризуют различные аспекты его функционирования, а различные классы соответствуют различным возможным стратегиям управления.

В настоящее время как зарубежными, так и отечественными исследователями накоплен большой опыт в использовании систем распознавания и разработано большое количество алгоритмов решения задач классификации. Однако внедрение этих алгоритмов в реальные системы управления по-прежнему затруднено рядом проблем, среди которых важное место занимает проблема нехватки и неточности экспертной информации об объектах управления [2]. Во многих случаях эксперты не обладают достаточным количеством информации для принятия гарантированно безошибочных решений о классах представленных им для классификации объектов. Это не позволяет построить систему распознавания, принимающую решения с необходимой точностью.

* Поддержано грантом РФФИ-Урал № 01-01-96419 и грантом МО № Т00 / 3.2 / 2647.

В такой ситуации возникает потребность в переходе к новой методологии взаимодействия экспертов и системы распознавания [1]. Она должна обеспечить использование возможностей методов распознавания образов по анализу многомерных данных, с одной стороны, и способностей экспертов, с другой стороны, как двух взаимосвязанных и активно взаимодействующих элементов в системе поддержки принятия решений. В ходе такого взаимодействия появляется возможность целенаправленного уточнения и корректировки априорных сведений за счет «согласования позиций» экспертов и системы распознавания. Разработка формальных процедур и математических методов, позволяющих организовать такое взаимодействие, является весьма важной и актуальной задачей.

Постановка задачи

В стандартной постановке задач распознавания образов различают два случая: дискриминантный и кластерный анализ [3]. В первом случае ставится задача обучения системы классификации: имеется некоторый набор объектов (обучающая выборка), заданных значениями своих измеряемых характеристик. Для каждого объекта априорно, на основании суждений экспертов, задан его класс. Необходимо, опираясь на эти данные, построить решающее правило, т.е. правило отнесения произвольного объекта с известными значениями измеряемых параметров к тому или иному классу. Во втором случае – при кластерном анализе – ставится задача самообучения. Она отличается тем, что для объектов обучающей выборки значения их принадлежности к классам заранее не известны. При этом методы распознавания образов опираются на гипотезу о компактности, т.е. предположение о том, что объекты одного и того же класса имеют близкие, в некотором смысле, характеристики, а объекты разных классов существенно различаются с точки зрения значений своих параметров.

В условиях неопределенности априорной информации о классах объектов применение методов первого типа оказывается неэффективным, поскольку для обучения используется априорная информация, содержащая значительную долю ошибок. С другой стороны, отказ от использования имеющихся неточных экспертных суждений, т.е. применение методов самообучения, существенно снижает возможности системы классификации по анализу данных, т.к. происходит потеря значительной части априорной информации и замена ее на искусственно введенные правила определения близости объектов. Поэтому наиболее эффективный подход в данном случае – рассматривать обучение и самообучение как единую задачу извлечения максимума информации из имеющихся данных.

При этом можно организовать итерационный процесс адаптации системы распознавания к ненадежным источникам априорной информации, т.е. к экспертам, выносящим свои суждения, и методам измерения параметров анализируемых объектов. Тогда процесс работы системы распознавания можно представить следующим образом.

Априорная экспертная классификация предъявляется системе распознавания для обучения и построения решающего правила. Система распознавания должна определить достоверность предоставленной информации для каждого объекта (на основе методов самообучения) и построить (на основе методов обучения) решающее

правило с учетом неодинаковой надежности и противоречивости различных элементов входной информации. Построенное в результате применения такой процедуры решающее правило будет классифицировать часть объектов иным образом, чем это было определено экспертами.

Далее вновь полученная классификация объектов предьявляется экспертам. С частью из внесенных в нее изменений эксперты могут не согласиться. В результате системе распознавания может быть предьявлена новая (исправленная) классификация. Таким образом, можно организовать процесс обмена классификациями между экспертами и системой распознавания. В ходе этого процесса решающее правило последовательно приближается к истинной неизвестной зависимости классов объектов от значений их параметров. Соответственно, качество работы системы распознавания при классификации новых (не входивших в обучающую выборку) объектов возрастает.

В рамках данной работы задача заключается в создании численных процедур, позволяющих организовать рассмотренный объединенный процесс обучения и самообучения.

Выбор решающих правил

Для решения поставленной задачи воспользуемся комитетными решающими правилами (комитетами большинства) [4, 5].

Определение [5]. Комитетом большинства системы линейных неравенств

$$\langle c; \mathcal{X}^j \rangle < 0, \quad j = \overline{1, m}, \quad (1)$$

где $\mathcal{X}^j \subset R^n$, $c \in R^n$ называется такое множество $\{c^1, \dots, c^q\} \subset R^n$, что каждому неравенству системы удовлетворяет более половины элементов этого множества (членов комитета).

Будем считать, что система (1) содержит $(n+1)$ неизвестных. Разложим каждый член комитета на составляющие $c^i = (a^i, b^i) = (a_1^i, \dots, a_n^i, b^i) \in R^{n+1}$, $i = \overline{1, q}$, и выберем коэффициенты неравенств специального вида:

$$\begin{cases} \mathcal{X}^j = (x^j, -1), \quad \forall x^j \in A, \\ \mathcal{X}^j = (-x^j, 1), \quad \forall x^j \in B, \end{cases} \quad j = \overline{1, m},$$

где $A = \{x^v\}$, $B = \{x^w\}$, $A \cup B = \{x^j, j = \overline{1, m}\}$, $A \cap B = \emptyset$ – обучающая выборка, состоящая из объектов двух классов (A и B). Векторы a^i представляют собой нормальные векторы гиперплоскостей и вместе с числами b^i задают q гиперплоскостей в пространстве параметров объектов. Векторы x^j определяют соответствующие объектам точки в том же пространстве. Разделяющий комитет $\{c^1, \dots, c^q\}$ можно рассматривать как решающую функцию, поскольку он обладает следующим свойством:

$$\begin{cases} \sum_{i=1}^q [\langle a^i; x \rangle - b^i < 0] > q/2 \quad \forall x \in A, \\ \sum_{i=1}^q [\langle a^i; x \rangle - b^i < 0] \leq q/2 \quad \forall x \in B. \end{cases}$$

Для решения поставленной задачи извлечения информации из обучающей выборки необходимо модифицировать понятие комитета. Как было указано выше, решающее правило должно классифицировать объекты так же, как в априорной классификации, если это не противоречит положению других объектов обучающей выборки. Решение комитета большинства можно рассматривать как результат голосования отдельных членов комитета (гиперплоскостей) «за» или «против» принадлежности объекта тому или иному классу. Т.е. если задан комитет, то для любого объекта может быть определен не только его класс, но и ранг (степень уверенности комитета в принятом решении):

$$r = \begin{cases} \sum_{i=1}^q [\langle a^i; x \rangle - b^i < 0] , & \text{если } x \in A, \\ \sum_{i=1}^q [\langle a^i; x \rangle - b^i > 0] , & \text{если } x \in B. \end{cases} \quad (2)$$

Тогда задачу извлечения информации из обучающей выборки можно рассматривать как задачу построения комитета, обеспечивающего максимизацию некоторой функции от рангов объектов. В качестве такой функции можно выбрать, например, суммарный ранг всех объектов обучающей выборки. Необходимость предлагаемой модификации обусловлена двумя причинами:

1) реально существующие задачи распознавания и классификации, в силу указанных выше особенностей, как правило, не позволяют четко разграничить случаи обучения и самообучения системы распознавания;

2) для комитетных решающих правил стандартного вида, несмотря на их высокую эффективность в задачах большой размерности, пока не найдено эффективных алгоритмов обучения, сочетающих достоинства малой вычислительной сложности и высокой степени обобщения информации.

Алгоритм построения комитета

Для построения комитета, обладающего указанными свойствами, необходимо организовать итерационный процесс. На каждом шаге этого процесса положение гиперплоскостей, составляющих комитет, должно изменяться так, чтобы обеспечить возрастание, или хотя бы неубывание, значения целевой функции

$$R = \sum_{i=1}^m r_i ,$$

где r_i – ранг i -го объекта обучающей выборки, вычисленный по формуле (2), m – размер обучающей выборки.

Пусть на некотором шаге получен комитет $\{c^1, \dots, c^q\} \subset R^{n+1}$. Вычислим ранги всех точек обучающей выборки по формуле (2). Для какой-нибудь из q гиперплоскостей, составляющих комитет, найдем объект $x^v \in A \cup B$, правильно классифицированный ею (с точки зрения априорной классификации), и объект $x^w \in A \cup B$, классифицированный данной гиперплоскостью неправильно. При этом необходимо, чтобы выполнялось условие:

$$r_v \leq r_w . \quad (3)$$

Стоит подчеркнуть, что речь идет о правильности классификации объектов именно отдельной гиперплоскостью, а не комитетом в целом. Пусть $I \subset \{1, \dots, m\}$ – множество номеров объектов обучающей выборки, правильно классифицированных данной гиперплоскостью. Согласно введенным обозначениям $v \in I$, $w \notin I$. Попытаемся изменить положение гиперплоскости так, чтобы обеспечить $w \in I$, а требование $v \in I$ снимем. Если это удалось, то новое множество номеров правильно классифицированных объектов будет иметь вид $I' \supset \{w\} \cup I \setminus \{v\}$ или $I' \supset \{w\} \cup I$, в зависимости от того, удалось ли сохранить принадлежность объекта x^v к числу правильно классифицированных. При этом новые значения рангов объектов:

$$\begin{aligned} r'_i &= r_i + 1, \quad i = w, \\ r'_i &\geq r_i - 1, \quad i = v, \\ r'_i &= r_i, \quad i \in I \setminus \{v\}, \\ r'_i &\geq r_i, \quad i \in \{1, \dots, m\} \setminus [I \cup \{w\}], \end{aligned} \quad (4)$$

т.е. новое значение суммарного ранга всех точек по крайней мере не меньше его старого значения. Если же при изменении положения гиперплоскости удастся сохранить правильную классификацию объекта x^v или добавить к числу правильно классифицированных какие-либо из объектов x^i , $i \notin I \cup \{w\}$, то значение целевой функции увеличивается.

Таким образом, задавшись некоторым начальным положением комитета, можно организовать итерационный процесс коррекции положения гиперплоскостей, ведущий к увеличению значения целевой функции. Рассмотрим функцию следующего вида:

$$R^0 = \sum_{i=1}^m \left[\left(\sum_{j=1}^q s_{ij} \right)^2 \right],$$

где $s_{ij} = 1$, если j -я гиперплоскость правильно классифицирует i -й объект и $s_{ij} = 0$ – в противном случае. Согласно выражениям (4), с учетом требования (3), значение функции R^0 строго возрастает на каждом шаге итерационного процесса. Поскольку существует точная верхняя граница значения этой функции:

$$R^0 = \sum_{i=1}^m \left[\left(\sum_{j=1}^q s_{ij} \right)^2 \right], \quad s_{ij} \in \{0, 1\} \Rightarrow R^0 \leq mq^2,$$

процесс завершится за конечное число шагов.

Если пересчитывать ранги объектов после перемещения каждой гиперплоскости, то изменение положения комитета будет в значительной степени зависеть от порядка выбора гиперплоскостей, подлежащих перемещению. Поэтому целесообразно на каждой итерации перемещать все плоскости, а уже потом рассчитывать новые значения рангов объектов. Исходя из вышесказанного, алгоритм построения комитета, извлекающего максимум информации из обучающей выборки, должен выглядеть следующим образом.

1. Задать начальное положение комитета случайным образом.
2. Рассчитать ранги всех объектов обучающей выборки по формуле (2).
3. Выбрать для обработки очередную гиперплоскость.
4. Попытаться найти еще не рассмотренную пару правильно и неправильно классифицированных ею объектов, удовлетворяющих условию (3).
5. Если объекты с требуемыми свойствами не найдены, то перейти к шагу 7, иначе – попытаться изменить положение гиперплоскости в соответствии с вышеописанными правилами.
6. Если положение гиперплоскости удалось изменить, то перейти к шагу 7, иначе – перейти к шагу 4.
7. Если все гиперплоскости уже рассмотрены, то перейти к шагу 8, иначе – перейти к шагу 3.
8. Если хотя бы одна из гиперплоскостей изменила свое положение, то перейти к шагу 2, иначе – искомым комитет найден.

В приведенном алгоритме имеется две операции, вычислительная сложность которых определяет аналогичный показатель для всего алгоритма. Это поиск объектов x^v и x^w – кандидатов на изменение рангов – и проверка возможности перемещения гиперплоскости, такого, что будут выполняться условия (4). Для поиска пар объектов x^v и x^w необходимо организовать перебор всех возможных сочетаний (v, w) , $v \in I$, $w \in \{1, \dots, m\} \setminus I$, удовлетворяющих условию (3). После того как была выбрана очередная пара (x^v, x^w) , необходимо найти такое положение гиперплоскости, которое обеспечит выполнение условий (4), либо определить, что у этой задачи решения нет. Геометрически это означает необходимость проверки линейной разделимости множеств:

$$\begin{aligned} & \mathcal{A}^0 = \{x^j : x^j \in A, j \in \{w\} \cup I \setminus \{v\}\} \\ \text{и} & \mathcal{B}^0 = \{x^j : x^j \in B, j \in \{w\} \cup I \setminus \{v\}\}. \end{aligned}$$

Т.е. нужно решить систему линейных неравенств:

$$\begin{cases} \langle x^j; a^i \rangle - b^i < 0, & x^j \in \mathcal{A}^0 \\ \langle x^j; a^i \rangle - b^i > 0, & x^j \in \mathcal{B}^0 \end{cases} \quad (5)$$

относительно $c^j = (a_1^j, \dots, a_n^j, b^j) \in \mathbb{R}^{n+1}$ либо определить, что решения нет. Для решения указанной задачи можно воспользоваться модификацией метода линейной коррекции. Он является эвристическим и, в принципе, не гарантирует нахождения решения системы (5). В то же время практическое его применение полностью оправдано в силу двух причин:

1. Временная сложность этого алгоритма линейно зависит от размерности задачи, т.е. он пригоден для использования в реальных задачах распознавания большой размерности.
2. Итерационный процесс поиска решения начинается от некоторого начального приближения. В качестве такого приближения можно выбрать текущее положение гиперплоскости. При очередном перемещении гиперплоскости в число правильно классифицированных ею должен войти только один объект – x^w). Поэтому большинство ограничений, налагаемых системой (5), выполняется уже для начального приближения. Т.е. обнаружение решения, если оно есть, следует ожидать уже после небольшого количества итераций.

Таким образом, максимальное количество итераций алгоритма линейной коррекции можно ограничить очень небольшим числом. Если решение после их выполнения не найдено, то можно прекратить поиск.

Пример использования алгоритма

Следует отметить, что разработанные методы не имеют жесткой привязки к какой-либо конкретной предметной области или к использованию каких-либо определенных параметров. Поэтому они могут быть использованы для решения задач принятия решений в любой прикладной области на различных уровнях управления. В частности, предложенные методы уже были использованы при построении системы распознавания для диагностики кардиологических заболеваний [6-8]. Ставилась задача оценивания тяжести заболевания. В качестве обучающей выборки рассматривалась группа из 75 больных. Каждый больной описывался вектором числовых и логических параметров, которые являются результатами инструментальных измерений и простых экспертных суждений, не требующих высокой квалификации медицинского персонала. Общее количество используемых параметров составило 120. Больные были разделены по степени тяжести заболевания на три группы: легкие, средние и тяжелые. Относительно некоторых больных (30 % от общего числа) экспертами-медиками был достоверно определен класс тяжести. Для остальных больных класс тяжести также задавался априорно, однако эксперты допускали возможность его ошибочного определения.

В результате автоматической классификации система распознавания изменила значения классов тяжести более чем у 40 % больных. В половине из этих случаев эксперты согласились с выводами системы распознавания. После второго обмена классификациями между экспертами и системой распознавания удалось довести степень их согласованности до 95 %. Для проверки обобщающей способности полученных решающих правил в рассмотрение было введено еще 25 больных с известными классами тяжести (тестовая выборка). В 90 % случаев система распознавания правильно определила тяжесть кардиомиопатии. Эта величина значительно превышает точность использовавшихся ранее для диагностики эвристических методов, позволявших правильно определить диагноз не более чем в 65 % случаев.

Заключение

1. Предложена новая методика построения системы распознавания, которая позволяет объединить процедуры обучения и самообучения в рамках единой процедуры автоматической классификации, обеспечивающей извлечение максимума информации из предоставленных данных и целенаправленное уточнение априорной информации.
2. В качестве основного элемента системы классификации использована модификация комитетных решающих правил, для которой предложен эффективный алгоритм обучения, использующий особенности постановки задачи.

3. Эффективность разработанной методики и алгоритмов проверена при решении практической задачи оценивания тяжести кардиологического заболевания. Полученные показатели качества работы системы распознавания (95 % – для решения единой задачи извлечения информации из обучающей выборки и 90 % – при классификации объектов тестовой выборки) говорят о высокой эффективности применения метода к решению разнообразных прикладных задач классификации.

Литература

1. Захаров В.Н. Интеллектуальные системы управления: основные понятия и определения // Известия РАН. Сер. Теория и системы управления. – 1997. – № 3. – С. 56-74.
2. Лапко А.В., Лапко В.А., Ченцов С.В. Непараметрические модели распознавания образов в условиях малых выборок // Автометрия. – 1999. – № 6. – С. 105-113.
3. Ту Дж., Гонсалес Р. Принципы распознавания образов: Пер. с англ. – М.: Мир, 1978. – 417 с.
4. Мазуров В.Д. Метод комитетов в задачах оптимизации и классификации. – М.: Наука, 1990. – 248 с.
5. Мазуров В.Д., Хачай М.Ю. Комитетные конструкции // Известия УрГУ. Сер. Математика и механика. – 1999. – № 14. – С. 77-108.
6. Кувшинов Б.М., Ширяев О.В., Шапошник И.И. Система диагностики заболеваний методами распознавания образов и классификации в n-мерном пространстве // Информационные технологии. – 2000. – № 6. – С. 43-47.
7. Кувшинов Б.М., Ширяев О.В., Богданов Д.В., Шапошник И.И., Ширяев В.И. Система классификации многопараметрических объектов для задач распознавания образов с неточной априорной информацией // Информационные технологии. – 2001. – № 11. – С. 35-43.
8. Кувшинов Б.М., Шапошник И.И., Ширяев В.И., Ширяев О.В. Использование комитетов в задачах распознавания образов с неточными экспертными оценками // Известия РАН. Сер. Теория и системы управления. – 2002. – № 5.

New approach to the problem of multivariable objects automated classification by the methods of pattern recognition is suggested that takes into account experts' knowledge lack during recognition system construction. It allows to combine usage of experts' judgements and data mining methods within a united process of the recognition system training. The basis of the approach lies in a new modification of the committee decision rules. Training algorithm is offered for decision rules search in a case of a priori information uncertainty. Effectiveness of the algorithms developed is tested while solving a practical medical problem of hypertrophy cardiomyopathy degree estimation.

Статья поступила в редакцию 16.07.02.