Предобработка данных с учетом заданных значений отдельных признаков

И.А. Евдокимов, В.Н. Гридин, В.И. Солодовников, И.В. Солодовников

Аннотация. Одной из ситуаций в значительной степени, усложняющей процесс распознавания образов, является ситуация, когда объекты из одного класса попадают в разные несвязанные между собой области. Поэтому в работе и ставится задача выявления таких областей (подклассов для объектов каждого класса).

Ключевые слова: предобработка данных, классификация, многосвязные области, класс, подкласс, мера близости

Введение

С ростом объемов хранимых данных особое значение приобретают вопросы предобработки данных. Исследователь испытывает естественные затруднения при выборе конкретного метода ввиду необозримости получаемой им информации. Во многих случаях для этого необходимо знать расположение объектов в пространстве, их взаимные зависимости и т.п.

При решении задачи классификации, например, полезно знать, насколько компактно размещены объекты одного класса. Одним из наиболее сложных случаев являются многосвязные области объектов, когда точки данного класса расположены в изолированных друг от друга областях. Во многих случаях представляет интерес разбиение классов на подклассы.

При анализе данных возникает задача распределения объектов в пространстве признаков в зависимости от фиксированных значений некоторых из них.

1. Постановка задачи

Пусть в пространстве признаков размерности M задано множество примеров/объектов.

$$O = \{O_1, \cdots, O_N\}.$$

Для пространства признаков установлено множество признаков:

$$P = \{x_1, \ldots, x_n, y_1, \ldots, y_m\},\ M = n + m$$

Каждый объект характеризуется множеством значений этих признаков

$$O_i = \{x_1^i, \ldots, x_n^i, y_1^i, \ldots, y_i^m\}$$

Признаки подразделяются на два подмножества:

$$X = \{x_1, \ldots, x_n\}$$

$$Y = \{y_1, \ldots, y_m\}$$

С помощью первого подмножества будет определяться принадлежность объектов к тому или иному подклассу. Второе подмножество задает набор ограничений или дополнительный критерий отбора. Предполагается, что два объекта могут принадлежать одному подклассу тогда и только тогда, когда у них совпадают все значения из множества Y:

$$\forall O_i, O_j\,(i \neq j) \wedge \left(O_i \in SubClass(K)\right) \wedge$$

$$\wedge \left(O_j \in SubClass(K)\right) \Rightarrow \forall k\,(k = 1 \div m)\ \left(y_k^i = y_k^j\right).$$

Здесь *SubClass* – функция, сопоставляющая объект с подклассом. Таким образом, под подклассом будем понимать объекты, подобные по подмножеству признаков X и эквивалентные по подмножеству признаков Y. В этом случае комбинация значений признаков Y может рассматриваться как своего рода метка класса [2]. При

интерпретации результатов эта метка может быть сопоставлена с конкретным классом.

Если исследователю известны классы объектов, то они могут быть введены в пространство признаков как единственный элемент подмножества *Y*. В этом случае определяются подклассы заранее определенных классов.

Будем предполагать, что пространство метрическое, либо координаты объектов являются результатом ранжирования, либо, по крайней мере, получены из матрицы различий согласно принципу метрического шкалирования.

В качестве меры близости будем рассматривать меру, удовлетворяющую условиям, накладываемым на расстояния. Очевидно, что результаты разбиения на подклассы при выборе одной могут отличаться от результатов, полученных для другой меры, но выбор меры во многом определяется целью задачи и знаниями о характере данных. Поэтому иногда целесообразно проводить исследование для различных мер близости.

Выделение подклассов происходит в пространстве размерности *n*, то есть на основании значений признаков из первого подмножества *X*.

Таким образом, задача формулируется в следующем виде:

Задано множество объектов $O = \{O_1, \cdots, O_N\}$.

Требуется построить множество подклассов $SC = \{SC_k\}$ такое, что расстояние между объектами в пространстве признаков *X* минимально, а в пространстве признаков *Y* равно нулю [1].

2. Алгоритм

Рассмотрим алгоритм, решающий данную задачу.

Начальный этап. Выбрать первый объект $(i = 1)$ из имеющейся выборки. Установить k=1 (номер кластера) и поместить объект в этот кластер. Задать начальное расстояние $d_0 = 0$.

Основная часть. Выполнять последующие действия для каждого объекта выборки.

А. Определить ближайший объект выборки, для которого $d_m \geq d_0$, d_m - расстояние до этого объекта. При этом возможны четыре варианта:

1. Ближайшим оказывается объект, принадлежащий к другому классу. В этом случае, формирование подкласса завершается. $k = k + 1$, $i = i + 1$. $d_0 = 0$. Перейти к А.

2. Ближайшим оказывается объект, принадлежащий к тому же классу, но не принадлежащий ни к какому подклассу. Этот объект включается в текущий подкласс. $d_0 = d_m$. Перейти к А.

3. Ближайшим оказывается объект, принадлежащий к тому же классу, но уже входящий в какой-либо другой подкласс. Два подкласса объединяются. Объединенному подклассу присваивается номер k. $d_0 = d_m$. Перейти к А.

4. Ближайшим оказывается объект, принадлежащий к тому же классу и входящий в тот же подкласс. ($d_0 = d_m$). Перейти к А.

3. Особенности

Алгоритм в некотором смысле подобен агломеративным алгоритмам кластерного анализа [3]. Однако в нем не возникает проблемы количества подклассов. Фактически это количество определяется на основании условий 1, 3 алгоритма. Поэтому условно его можно отнести к алгоритмам контролируемой кластеризации [4], в которых допустимо вводить дополнительные условия (даже «извне») на правила отбора объектов в кластер.

Даже если точки с одной меткой класса находятся в односвязной области, класс может быть разбит на несколько подклассов. Эту особенность алгоритма можно сформулировать следующим образом: если некоторый класс объектов, находящийся в связанной области, состоит из нескольких компактных групп объектов, достаточно удаленных друг от друга, и при этом каждая группа тяготеет к объектам другого класса, то этот класс будет разбит на количество подклассов, равное количеству групп объектов подобных объектам другого класса.

Пусть одно сочетание значений признаков подмножества *X* однозначно определяет значение признаков подмножества *Y*, то есть объекты различимы по *X*,

$$\{\forall i, j\}\left(\nu_i(x_1, \cdots, x_n) = \nu_j(x_1, \cdots, x_n) \Rightarrow i = j\right),$$

ν_i - вектор значений признаков для i-го объекта. В этом случае разбиение на подклассы не зависит от порядка предъявления объектов выборки.

В противном случае разбиение на подклассы будет зависеть от порядка предъявления объектов. При этом для одного из таких объектов может образоваться подкласс, состоящий только из этого объекта.

При предъявлении нового объекта возможно возникновение следующих ситуаций:

1. Метка класса добавляемого объекта не встречалась, тогда формируется новый подкласс, состоящий из одного объекта.

2. Ближайшим к добавляемому объекту является объект с той же меткой класса. В этом случае новый объект присоединяется к подклассу ближайшего объекта

3. Ближайшим к добавляемому объекту является объект с меткой, отличной от его метки. В этом случае формируется новый подкласс, состоящий из одного включаемого объекта.

Наконец, в качестве критерия отнесения к подклассу для множества признаков Y можно использовать не совпадение значений признаков, а допустить их различие на некоторую $\delta \to 0$.

4. Пример

Рассмотрим пример, демонстрирующий работу алгоритма.

Пусть метка класса может принимать три значения (условно, обозначим их 1, 2, 3), а исходная выборка имеет следующий вид (Рис. 1). Здесь также предполагается, что объекты со значением метки класса 1 занимают три не связанных между собой области пространства, объекты с меткой, имеющей значение 2, занимают одну область, но с достаточно сложной границей, а объекты со значением метки класса 3 – две области пространства, имеющих сложные границы.

На Рис.2 приведено разбиение объектов на подклассы с использованием расстояния по Евклиду. В этом случае объекты с меткой 1 не разбиваются на подклассы. Объекты с меткой 3 в одной из областей разбиваются на три подкласса, а для объектов с меткой 2 выделяется 5 подклассов. Здесь границы подклассов проведены приблизительно (в настоящий момент программа не строит границы).

Рассмотрим образование подкласса 2.2. Расстояние d_1 от объекта, входящее в него, меньше до объекта, входящего в подкласс 3.4, чем до

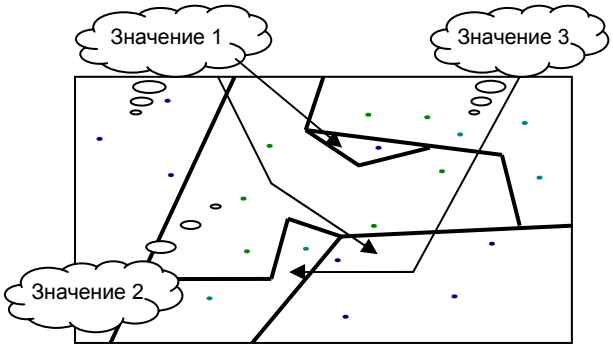

Рис. 1, Исходные данные

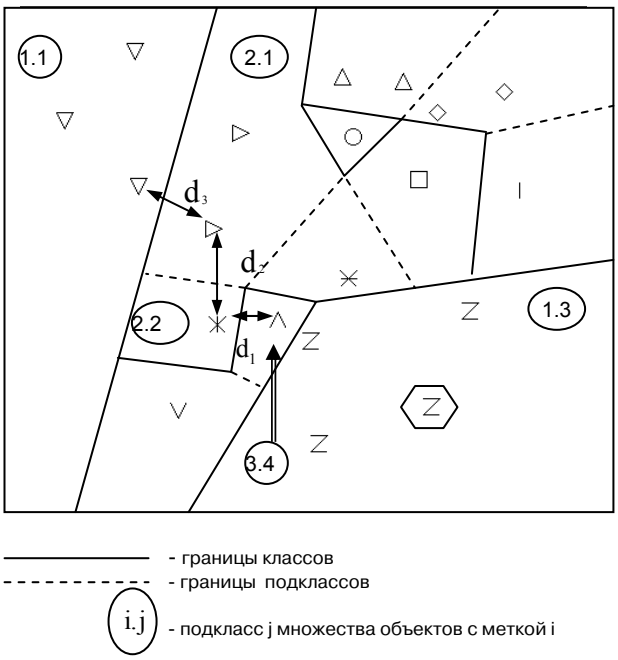

Рис. 2. Разбиение на подклассы с использование расстояния по Евклиду

объекта, имеющего ту же метку ($d_1 < d_2$). С другой стороны, расстояние до ближайшего объекта с той же меткой больше, чем расстояние от этого объекта до объекта из подкласса 1.1 ($d_2 < d_3$).

Несколько иная картина возникает для объектов, оказавшихся в подмножестве 1.3. Здесь один из объектов явно ближе к объектам с метками 3 и 2, но, тем не менее, для некого не выделяется отдельного подкласса. В этом случае ключевым оказывается объект, заключенный в шестиугольник. Дело в том, что расстояние от него до любого объекта области 1.3, меньше чем расстояние до любых объектов из других областей.

Заключение

Алгоритм использовался для предобработки данных в целях формирования одной из парадигм нейронных сетей LVQ-сети, являющейся развитием сетей Кохонена, которая может использоваться как для классификации, так и для кластеризации, а также для решения ряда социометрических задач.

Литература

1. Евдокимов И.А., Серегина Ю.А., Солодовников И.В. Формирование LVQ-сети для задачи классификации – Материалы Девятого научно – практического семинара «Новые информационные технологии», Москва, 2007
2. Комарцова Л.Г., Максимов А.В. Нейрокомпьютеры: Учеб. Пособие для вузов – М.: Изд-во МГТУ им. Н.Э. Баумана, 2004
3. Барсегян А.А. Технология анализа данных: Data Mining, Visual Mining, Text Mining, OLAP – СПб.: БХВ – Петербург, 2007
4. Рассел С., Норвиг П. Искусственный интеллект: современный подход – М.: Издательский дом «Вильямс», 2006.

Евдокимов И.А. Бакалавр, магистратура МГТУ им. Н.Э. Баумана, 2 публикации.

Гридин В.Н. Директор ЦИТП РАН, доктор технических наук, профессор, заслуженный деятель науки и техники, 169 публикаций. Эл. адрес: info@ditc.ras.ru.

Солодовников В.И. Научный сотрудник ЦИТП РАН, кандидат технических наук, 17 публикаций.

Солодовников И.В. Главный научный сотрудник ЦИТП РАН, доктор технических наук, профессор МГТУ им. Н.Э. Баумана, 88 публикаций. Тел. 235-83-82.