

Метод распознавания образов на основе использования критерия Стьюдента

В.Г.РОДЧЕНКО

Введение

При решении целого ряда прикладных задач в области технических, гуманитарных, социально-экономических дисциплин приходится сталкиваться с проблемой исследования объектов сложной природы, которые описываются большим числом разнообразных признаков. Использование традиционного математического аппарата оказывается в данном случае или весьма затруднительным, или даже невозможным [1]. При этом более эффективно проявили себя относительно новые подходы, связанные с применением методов прикладной статистики и, в частности, математической теории распознавания образов [2,3]. Процесс распознавания строится на основе анализа описывающих объекты признаков, задаваемых в виде классифицированной обучающей выборки. При этом объективность и достоверность распознавания обеспечивается реализацией специального предварительного этапа исследования, который связан с анализом возможности построения разделенных в многомерном признаковом пространстве эталонов классов.

Постановка задачи

Предположим, что имеется k классов, каждый из которых задается набором из m_i (где $i = \overline{1, k}$) многомерных объектов. При этом, во-первых, объекты описываются n числовыми признаками из априорного словаря, во-вторых, все выборки значений признаков каждого класса имеют нормальные распределения, в-третьих, для каждого объекта известно, к какому классу он принадлежит. Набор из всех объектов одного класса образует исходное описание этого класса в априорном признаковом пространстве. Объединение же всех объектов из всех классов будет представлять собой классифицированную обучающую выборку (КОВ). Эта выборка представляет собой таблицу типа "объект-свойство" и формально записывается в виде матрицы размерности $n \times m$, где $m = m_1 + m_2 + \dots + m_k$, а m_i – количество объектов i -го класса.

Процесс распознавания начинается с проведения разведочного анализа информативности признаков, включенных в априорный словарь. Для этого используется заданная классифицированная обучающая выборка. Признаки по степени информативности с точки зрения разделения классов в многомерном признаковом пространстве сепарируются на три вида.

К первому виду будут отнесены те признаки, для которых выполняется следующее условие: *для всех пар классов на основе использования критерия Стьюдента оказалось, что выборки значений этого признака для двух сравниваемых классов не показали существенного различия между ними.* Отметим, что признаки первого вида имеют значения, фактически распределенные по одному закону во всех без исключения классах, а значит, они не могут выступать в качестве "разделителей" классов в признаковом пространстве.

Если для очередного анализируемого признака выполняется условие: *для всех пар классов на основе использования критерия Стьюдента оказалось, что выборки значений этого признака для двух сравниваемых классов показали существенное различия между ними,* то этот признак относится ко второму виду. Именно такие признаки и будут разделять

образы классов в многомерном признаковом пространстве, а потому только они и включаются в рабочий словарь признаков (РСП), на основе которого будут описываться объекты при дальнейшем распознавании.

Признаки, для которых не выполняется ни первое, ни второе условие, будут относиться к признакам третьего вида. Отметим, что эти признаки не отражают какие-либо четко выраженные различия между классами, а потому будут создавать “шумы” при распознавании и ухудшать его качество.

После проведения разведочного анализа информативности признаков и формирования рабочего словаря выполняется второй этап процесса распознавания, который связан с аттестацией признаков, включенных в РСП. Для этого из классифицированной обучающей выборки исключаются строки, которые соответствуют признакам первого и третьего видов, т.е. КОВ переформируется на основе РСП. Затем каждый класс разбивается на два одинаковых по количеству объектов подкласса, и на основе первого подкласса строится эталон класса, а второй используется для проверки правильности распознавания. Далее первый и второй подклассы меняются ролями и алгоритм повторяется вновь. Если были получены удовлетворительные результаты при распознавании, то осуществляется переход к следующему этапу, а иначе необходимо провести дополнительный анализ содержимого рабочего словаря признаков.

На заключительном этапе проводится непосредственно процедура распознавания исследуемого класса, предварительно описанного на основе признаков из рабочего словаря. На основе значений признаков из классифицированной обучающей выборки строятся эталоны классов и проводится распознавание исследуемого класса. В результате он или относится к одному из исходных классов, или выделяется в самостоятельный класс – “джокер-класс”.

Алгоритм реализации метода

Алгоритм для реализации метода распознавания на основе использования критерия Стьюдента предполагает выполнение следующих шагов:

Шаг 1. Формируются алфавит классов $A = \{A_1, A_2, \dots, A_k\}$ и априорный словарь признаков (АСП) $P = \{P_1, P_2, \dots, P_n\}$. Класс A_i (где $i = \overline{1, n}$) изначально определяется совокупностью объектов, каждый из которых в свою очередь на основе признаков из АСП описывается в многомерном признаковом пространстве в виде вектора-столбца $x^T = (x_1, x_2, \dots, x_n)$, где $x_i \in R$ для $\forall i = \overline{1, n}$. Объединение таких векторов из всех классов образуют классифицированную обучающую выборку. Эта выборка представляется в виде прямоугольной таблицы типа "объект-свойство", содержащей n строк и m столбцов (где $m = m_1 + m_2 + \dots + m_k$, а m_i – количество объектов i -го класса). При этом для $\forall A_i \subset A$, где $i = \overline{1, k}$ получаем соответствующую матрицу X_i размерности $n \times m_i$, где m_i – число объектов i -го класса. Все выборки значений признаков каждого класса имеют нормальные распределения. Из объектов исследуемого класса формируется матрица X_{k+1} размерности $n \times m_{k+1}$, где m_{k+1} – число объектов исследуемого класса.

Шаг 2. Производится сепарирование признаков по степени их информативности на основе критерия Стьюдента. Последовательно, начиная с первого и до последнего, анализируются признаки из априорного словаря $P = \{P_1, P_2, \dots, P_n\}$. В результате, они разбиваются на три вида $P^{(1)} = \{P_1^{(1)}, P_2^{(1)}, \dots, P_{n_1}^{(1)}\}$, $P^{(2)} = \{P_1^{(2)}, P_2^{(2)}, \dots, P_{n_2}^{(2)}\}$, $P^{(3)} = \{P_1^{(3)}, P_2^{(3)}, \dots, P_{n_3}^{(3)}\}$, где $P = P^{(1)} \cup P^{(2)} \cup P^{(3)}$ и $n_1 + n_2 + n_3 = n$.

Отнесение очередного анализируемого признака P_i (где $i = \overline{1, n}$) к одному из трех видов производится по следующему правилу:

- если для всех пар классов критерий Стьюдента не показал существенного различия между выборками значений этого признака для двух сравниваемых классов, то P_i – признак первого вида;

- если для всех пар классов критерий Стьюдента показал существенное различие между выборками значений этого признака для двух сравниваемых классов, то P_i – признак второго вида;
- если для признака P_i не выполнилось ни одно из двух предыдущих условий, то он относится к третьему виду.

Только признаки второго вида $P^{(2)} = \{P_1^{(2)}, P_2^{(2)}, \dots, P_{n_2}^{(2)}\}$ включаются в рабочий словарь, который и будет использоваться для дальнейшего исследования. При этом отметим, что переход к следующему шагу алгоритма происходит только в том случае, когда РСП оказывается непустым, а иначе необходимо возвращаться к началу и формировать новый вариант априорного словаря.

Шаг 3. Проводится реформирование матриц $X_1, X_2, \dots, X_k, X_{k+1}$ путем исключения в каждой из них строк, содержащих значения признаков первого $P^{(1)}$ и третьего $P^{(3)}$ видов. В итоге получаются матрицы $Y_1, Y_2, \dots, Y_k, Y_{k+1}$ размерности $n_2 \times m_i$, в которых все соответствующие значения признаков нормируются к единичному интервалу по формуле $y_i = (x_i - x_{\min}) / (x_{\max} - x_{\min}) \forall i = \overline{1, k+1}$.

Шаг 4. Все матрицы множества $Y = \{Y_1, Y_2, \dots, Y_k\}$ разбивается на две части так, что $\forall i = \overline{1, k} Y_i = Y_i^{(1)} + Y_i^{(2)}$, $Y_i^{(1)}$ – матрица размерности $n_2 \times m_i^{(1)}$ (где $m_i^{(1)} = m_i / 2$ – количество объектов, включенных в матрицу $Y_i^{(1)}$), $Y_i^{(2)}$ – матрица размерности $n_2 \times m_i^{(2)}$ (где $m_i^{(2)} = m_i - m_i^{(1)}$ – количество объектов, включенных в матрицу $Y_i^{(2)}$). В результате, получаются два подмножества $Y^{(1)} = \{Y_1^{(1)}, Y_2^{(1)}, \dots, Y_k^{(1)}\}$ и $Y^{(2)} = \{Y_1^{(2)}, Y_2^{(2)}, \dots, Y_k^{(2)}\}$. На основе матриц $Y^{(1)} = \{Y_1^{(1)}, Y_2^{(1)}, \dots, Y_k^{(1)}\}$ строятся эталоны $E^{(1)} = \{E_1^{(1)}, E_2^{(1)}, \dots, E_k^{(1)}\}$ для каждого класса, где

$$E_i^{(1)T} = (e_{i1}^{(1)}, \dots, e_{in_2}^{(1)}) \text{ и } e_{ij}^{(1)} = \frac{1}{m_j^{(1)}} \sum_{i=1}^{m_j^{(1)}} y_i^{(1)}.$$

ошибочных классификаций Q . После чего проводится классификация объектов из множества $Y^{(2)} = \{Y_1^{(2)}, Y_2^{(2)}, \dots, Y_k^{(2)}\}$ и подсчитывается число ошибочных классификаций G . Для классификации вводится метрика (например, евклидово расстояние), и тогда классификация объекта из $Y_i^{(2)} \forall i = \overline{1, k}$ будет ошибочной в том случае, когда этот объект расположен ближе к эталону $E_j^{(1)}$ (где $j \neq i$), чем к эталону $E_i^{(1)}$.

Затем подмножества $Y^{(1)} = \{Y_1^{(1)}, Y_2^{(1)}, \dots, Y_k^{(1)}\}$ и $Y^{(2)} = \{Y_1^{(2)}, Y_2^{(2)}, \dots, Y_k^{(2)}\}$ меняются ролями, и $Y^{(2)} = \{Y_1^{(2)}, Y_2^{(2)}, \dots, Y_k^{(2)}\}$ используется для построения эталонов классов, а объекты из $Y^{(1)} = \{Y_1^{(1)}, Y_2^{(1)}, \dots, Y_k^{(1)}\}$ распознаются.

Если в результате $G \leq Q$, то аттестация прошла удовлетворительно и осуществляется переход к следующему шагу алгоритма, а иначе необходимо вернуться к шагу 2 данного алгоритма и уточнить РСП.

Шаг 5. Последовательно для каждого объекта из Y_{k+1} ищется ближайший из эталонов множества $E^{(1)} = \{E_1^{(1)}, E_2^{(1)}, \dots, E_k^{(1)}\}$, и объект относится к тому классу, которому соответствует этот эталон.

После этого принимается решение либо об отнесении исследуемого класса к одному из исходных, либо об его выделении в отдельный класс.

Заключение

В качестве результата предлагается оригинальный метод распознавания, который предусматривает, во-первых, сепарирование признаков по степени их информативности на основе критерия Стьюдента с целью построения разделенных эталонов классов и, во-вторых, выполнение обязательного этапа по аттестации используемых для распознавания признаков. Данный метод может использоваться при построении специализированных систем для рас-

познавания объектов сложной природы, описываемых набором вещественных признаков, имеющих нормальные распределения.

Abstract

The article deals with the method of pattern recognition, which is to be applied in studies of multidimensional objects of complex nature, the latter comprising the set of substantial signs having normal distribution rate. The proposed method deals with the Student-criterion-based separation of signs according to the degree of their information satiety rate.

Литература

1. Н.Г.Загоруйко, Прикладные методы анализа данных и знаний, Новосибирск: Изд-во Института математики, 1999.
2. В.И.Васильев, Принцип простоты в проблеме обучения распознаванию образов, Распознавание, классификация, прогноз. Математические методы и их применение, Вып.3, М.: Наука, 1992.
3. В.Г.Родченко, Об одном методе построения системы распознавания образов, Известия Гомельского государственного университета имени Ф.Скорины, 2002. — № 6. — С. 93–96.

Гродненский государственный
университет им. Янки Купалы

Поступило 10.04.03