УДК 004.42:519.24

М.П. Лазеева, А.В. Дерюшев

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКОГО НЕПАРАМЕТРИЧЕСКОГО МЕТОДА РАСПОЗНАВАНИЯ ОБРАЗОВ

При управлении сложными системами и процессами (техническими, биологическими, экономическими, социальными и т. п.) необходимо оперативно принимать наиболее рациональные решения. Зачастую для принятия такого рода решений требуется установить принадлежность исследуемого объекта к одному из классов (образов) – множеству объектов, объединенных общими свойствами. Обычно такого рода задачи решают чаще на интуитивном уровне специалисты в конкретной области знаний. Возможность математической постановки ранее не формализованных и решаемых на интуитивном уровне задач классификации обусловила важность и перспективность применения методов распознавания образов широким кругом специалистов: экономистами, математиками, инженерами, социологами, геологами, медиками и т. д.

Распознавание образов – научно-техническое направление, связанное с разработкой методов и построением систем (в том числе на базе ЭВМ) для установления принадлежности некоторого объекта к одному из заранее выделенных классу объектов (образу). Процесс распознавания основан на сопоставлении признаков, характеристик исследуемого объекта с признаками, характеристиками других известных объектов, в результате чего делается вывод о наиболее правдоподобном их сопоставлении.

Можно выделить следующие основные понятия:

- объект – предмет или явление, исследуемое в конкретной задаче и характеризующееся некоторым конечным числом признаков;
- фактор (признак) – количественное или качественное описание того или иного свойства исследуемого объекта;
- решающее правило распознавания – алгоритм, по которому методом обучения на основе анализа значений вероятностных характеристик признаков делается вывод о том, к какому классу принадлежит объект.

Задачу обучения распознаванию можно сформулировать следующим образом. Имеется некоторое число объектов, относящихся к двум различным классам. Необходимо на основе информации о них найти правило, с помощью которого можно классифицировать другие объекты, описанные той

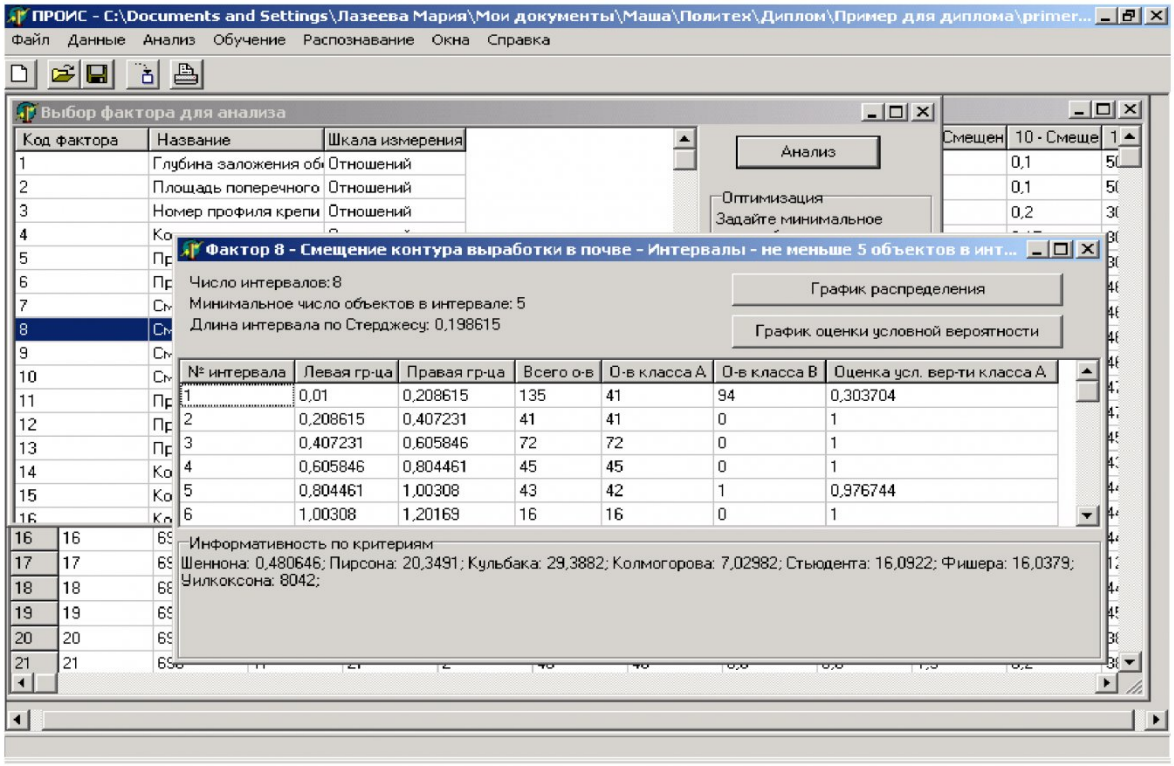

Рис. 1. Интерфейс ППП «ПРОИС» версии 2.0.0

же системой факторов, что и данные, при этом ошибки прогноза должны быть как можно меньшими. Исходное множество объектов будем называть обучающей выборкой. Чтобы осуществить обучение распознающей системы, необходимо произвести большой объем вычислений. Поэтому интерес к использованию методов распознавания возрос с появлением вычислительных машин с высоким быстродействием. Первоначально широкое распространение получили различные алгоритмы распознавания, основанные на методах математической статистики. Однако затем эти методы, на наш взгляд, были незаслуженно забыты с развитием нейроинформатики [1]. Распознавание на основе нейронных сетей позволяет эффективно решать ряд задач классификации. Но в большинстве случаев для решения конкретных задач необходимо создавать новую сеть, приспособленную для использования в данной области знания. Статистические методы более универсальны, многие из них могут быть использованы для решения задач классификации и прогноза в различных отраслях науки.

Нами был разработан пакет прикладных программ (ППП) «ПРОИС» версии 2.0.0, в котором реализован алгоритм вероятностно-статистического непараметрического метода распознавания образов [2]. Он предназначен для решения задач классификации и прогноза и позволяет разделить незнакомые объекты на два класса. Пакет предоставляет возможности для статистического анализа исходных данных, а также позволяет автоматизировать расчеты при обучении распознающей системы и непосредственно при распознавании.

Для работы с ППП «ПРОИС» необходимо сформировать обучающую выборку объектов с эффективной для распознавания системой факторов. Для ввода данных предназначены электронные таблицы факторов и выборки. Производится автоматическая проверка вводимых значений.

Основные возможности ППП «ПРОИС» 2.0.0 (рис. 1).

Вычисление основных точечных статистик выборки. К ним относятся:

- минимальное и максимальное значение фактора;
- математическое ожидание;
- дисперсия;
- среднеквадратическое отклонение;
- коэффициент вариации.

Определяются и некоторые другие характеристики анализируемой выборки: объем, число объектов первого и второго классов, а также оценка условной вероятности первого класса.

Построение эмпирической функции распределения и оценка информативности факторов. Для каждого фактора производится построение эмпирической функции распределения. ППП «ПРОИС» также предоставляет возможность графического анализа распределения объектов по интервалам.

Оценка информативности проводится по семи критериям:

- J – критерий Шеннона;
- χ^2 –критерий Пирсона;
- λ – критерий Колмогорова;
- I – критерий Кульбака;
- t – критерий Стьюдента;
- F – критерий Фишера;
- U – критерий Уилкоксона.

Анализ информативности факторов можно провести с помощью обобщенной таблицы информативности, а также ранжирования факторов методом обобщенной ранжировки с учетом мнений всех экспертов.

По результатам анализа обучающей выборки производится **минимизация признакового пространства** – формирование окончательной совокупности факторов, которая будет использована в процессе обучения. Исключаются наименее информативные факторы. Большую роль при минимизации признакового пространства играет профессиональный анализ факторов экспертами.

Обучение распознающего аппарата. Поскольку факторы имеют различную размерность и пределы измерений, могут иметь различные знаки, шкалы измерений, то объекты рассеяны во всевозможных октантах гиперпространства, компактность образов мала.

Для повышения эффективности обучения необходимо ввести безразмерные координаты. В качестве новой безразмерной координаты может быть использована оценка условной вероятности первого из классов. Для перевода значений факторов в безразмерные координаты производится построение таблиц перевода, которые содержат в себе границы изменения значений факторов по интервалам, и оценки условной вероятности первого из классов, соответствующие каждому интервалу.

На рис. 2 дана геометрическая интерпретация

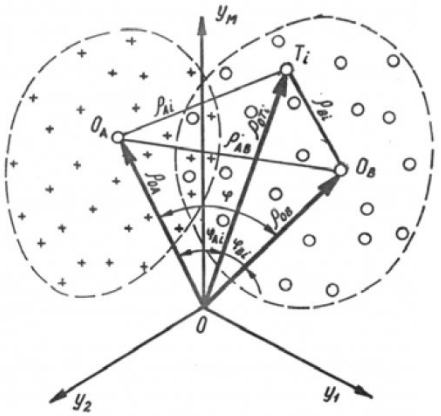

Рис. 2. Геометрическая интерпретация положения образов в гиперпространстве

положения образов в трехмерном пространстве. Можно определить некоторые характеристики положения образов в гиперпространстве, например: координаты центров тяжести образов, расстояние между центрами тяжести образов, расстояние от начала координат до центров тяжести образов, угол между радиус-векторами центров тяжести образов и т.д.

На основе этих характеристик производится разработка восьми решающих правил распознавания:

- по разности расстояний от объекта до центра тяжести образов;
- по разности углов между радиус-векторами объектов и центров тяжести образов;
- по разности скалярных произведений радиус-векторов объектов и центров тяжести образов;
- по разности между коэффициентами корреляции объектов и центрами тяжести образов;
- по расстоянию от объектов до гиперплоскости, проходящей по нормали к середине отрезка, соединяющего центры тяжести образов;
- по расстоянию от начала координат до каждого объекта;
- с помощью разделяющего гиперэллипсоида;
- по обобщенной функции желательности.

По каждому из решающих правил производится построение прогнозной таблицы, которая содержит границы изменения обобщенного признака по интервалам, а также оценки условной вероятности принадлежности объекта к первому из классов, соответствующие каждому интервалу.

Оценка ошибок распознавания. Для оценки ошибок распознавания используется контрольная выборка – множество объектов, описываемых окончательной системой факторов, которые не были включены в обучающую выборку. Ввод контрольной выборки производится аналогично вводу обучающей выборки. Происходит распознавание по каждому решающему правилу и оценка ошибок прогноза: ошибки первого рода – когда объект первого из классов классифицируется как объект второго; ошибки второго рода – когда объект второго из классов классифицируется как объект первого; а также суммарные ошибки.

При этом распознавание контрольной выборки можно провести с использованием различных порогов принятия решения – минимальных значений оценки условной вероятности, при которой объект будет отнесен к первому из классов.

В результате анализа ошибок распознавания определяют **оптимальное решающее правило распознавания** и **оптимальный порог принятия решения**. На этом обучение распознающего аппарата можно считать завершенным. Выбранные оптимальный порог принятия решения и оптимальное решающее правило будут использоваться для последующего распознавания объектов.

Распознавание. Объекты, подлежащие классификации, составляют экзаменационную выборку. Ее заполняют аналогично обучающей и контрольной выборкам. Результатом распознавания будет оценка условной вероятности принадлежности каждого объекта к первому из классов и класс объекта, определенный при установленном пороге принятия решения.

Разработанная система применима для решения задач классификации и прогноза в различных областях знаний и является ядром комплекса, предназначенного для статистической обработки данных и распознавания образов.

Данный комплекс предполагается использовать в Кузбасском государственном техническом университете на кафедре строительства подземных сооружений и шахт для прогнозирования надежности вскрывающих горных выработок.

СПИСОК ЛИТЕРАТУРЫ

1. *Осовский С.* Нейронные сети для обработки информации/ пер. с польского И.Д. Рудинского. – М.: Финансы и статистика, 2002. 344 с.

2. *Мякишева Л. Е.* Использование ЭВМ для решения задач методами распознавания образов: Методические указания по применению библиотеки научных программ "ПРОИС" / Л. Е. Мякишева, А. В. Дерюшев; КузНИИшахтострой; Кузбас. политехн. ин-т. – Кемерово, 1986. 40 с.

Авторы статьи:

Лазеева
Мария Петровна
– дипломант каф. вычислительной техники и информационных технологий (гр. ИС991)

Дерюшев
Александр Владимирович
- канд. техн. наук, доц. каф. строительства подземных сооружений и шахт