УДК 004:62-52:004.03

ИСПОЛЬЗОВАНИЕ КРИТЕРИЯ СТЬЮДЕНТА ДЛЯ ВЫЯВЛЕНИЯ СТАТИСТИЧЕСКИ ЗНАЧИМЫХ РАЗЛИЧИЙ В ЗАДАЧАХ КЛАССИФИКАЦИИ

К. т. н. Н. О. Комлевая, А. Ю. Махиненко, В. А. Луговской, В. В. Провоторов

Одесский национальный политехнический университет
Украина, г. Одесса
nokoml@yandex.ua

Рассмотрен подход, позволяющий использовать статистические методы для классификации объектов. Применение критерия Стьюдента позволило сравнить средние значения признаков в классах. Даны рекомендации для принятия решения о принадлежности объекта определенному классу. Предложено дальнейшее развитие работы с использованием статистических классификаторов, позволяющих уменьшить число параметров классификации.

Ключевые слова: критерий Стьюдента, классификация, статистика.

Задачей классификации часто называют предсказание категориальной зависимой переменной на основе выборки непрерывных и/или категориальных переменных. Процесс классификации заключается в разбиении множества объектов на классы по определенному критерию. Классификатором называется некая сущность, определяющая, какому из предопределенных классов принадлежит объект по вектору признаков. Для классификации используются различные методы [1], каждый из которых имеет свои преимущества. Основные из них: классификация с помощью деревьев решений; байесовская классификация; классификация при помощи искусственных нейронных сетей; классификация методом опорных векторов; статистические методы, в частности, линейная регрессия; классификация при помощи метода ближайшего соседа; классификация CBR-методом; классификация при помощи генетических алгоритмов.

Целью работы является рассмотрение возможности применения подхода для выявления статистически значимых различий между двумя группами (классами) с использованием критерия Стьюдента и выработка рекомендаций для принятия решения о принадлежности объекта тому или иному классу.

В качестве материала для исследований взяты две группы объектов, каждый из которых характеризуется вектором из 32 признаков [2]. Априори известно, что объекты классифицированы верно, при этом каждый класс содержит разное количество объектов. Также известно, что значения признаков являются количественными величинами, измеренными на непрерывной шкале.

Интерес представляет анализ и сравнение одноименных признаков, характеризующих объекты разных классов. Первым этапом анализа количественных данных является анализ вида их распределения. Существует несколько способов решения этой задачи. Для получения результата на качественном уровне достаточно построить график функции распределения и визуально оценить, насколько он близок к колоколу нормального распределения. Однако решение поставленной нами задачи требует числовых результатов. Одним из количественных способов является оценка симметричности распределения признаков, имеющих только положительные значения. Однако этот способ не позволяет оценить эксцесс. Поэтому для получения надежной оценки соответствия изучаемого распределения признака закону нормального распределения следует проверить статистическую гипотезу о виде распределения, от которого зависит выбор методов дальнейшего анализа данных.

Для принятия решения о виде распределения обычно используется один из следующих критериев: Колмогорова—Смирнова (для случаев, когда среднее значение и среднее квадратическое отклонение не вычисляются по выборке, а известны заранее), Лиллиефорса или Шапиро—Уилка (для случаев, когда отклонения признака заранее неизвестны). Если с учетом критерия распределение исследуемого признака можно считать нормальным, то при сравнении групп по этому признаку можно

пользоваться параметрическими методами, которые обладают большей статистической мощностью, чем непараметрические.

Для сравнения двух независимых групп по количественным признакам нами был использован параметрический метод — *t*-критерий Стьюдента. Данный метод заключается в проверке нулевой гипотезы о том, что средние значения признака в сравниваемых группах не различаются. Если нулевая гипотеза по результатам теста отклоняется, то следует принять альтернативную гипотезу о том, что средние значения в группах различны. Однако для применения этого метода в его классическом варианте требуется, чтобы дисперсии распределений признаков в двух сравниваемых группах были равны. Для проверки равенства дисперсий использовался метод Левена, с помощью которого определялись близость дисперсий одноименных признаков для объектов разных классов. Вычисленное *p*-значение теста Левена превысило критическое значение 0,05 для всех пар рассматриваемых признаков, что говорит о гомогенности дисперсий.

Для вычисления *t*-критерия Стьюдента была использована зависимость для неравночисленных выборок с учетом несмещенных оценок дисперсий [3]. Полученное $t_{эмп}$ сравнивали с критическим значением этого показателя $t_{кр}$, выбранным из стандартных таблиц по значениям k и p (при выбранном уровне статистической значимости $p > 0{,}05$). При $t_{эмп} > t_{кр}$ гипотеза H_0 о сходстве значений по соответствующим признакам из разных классов отклонялась и принималась альтернативная гипотеза H_1 о различии значений. Таким образом, для принятия решения о принадлежности объекта определенному классу должна приниматься гипотеза H_0 о сходстве значений между классом и исследуемым объектом по всем 32 элементам вектора признаков. Если же данное условие не может быть выполнено, то объект считается неклассифицированным. Данные о таких объектах целесообразно накапливать для дальнейших исследований.

Анализ результатов классификации показал, что различные признаки вносят различный вклад в процесс классификации, так как чем больше значение $t_{эмп}$ при выполнении условия $t_{эмп} > t_{кр}$, тем более значимым является данный признак. Поэтому в дальнейшем предлагается сформировать статистический классификатор – вектор $V=<V_1, V_2, \ldots, V_k>$, каждый элемент V_i (i=1..k) которого представляет собой пару ($Name_i$, St_i), где $Name_i$ – имя соответствующего признака, $St_i = t_{эмп}/t_{кр}$ для данного признака, St_i >1. Сортировка такого вектора по убыванию значений St_i должна привести к тому, что в начале вектора сосредоточатся признаки, в наибольшей степени учитывающие особенности различий между классами. Дальнейшие исследования должны показать минимально необходимое количество признаков, взятых с начала такого вектора, значения которых позволят различать классы при определенном уровне статистической значимости. Это должно привести к уменьшению числа параметров классификации.

Таким образом, описанный подход с применением критерия Стьюдента позволяет решить задачу классификации с использованием статистических методов. При необходимости сравнения между собой по количественному признаку трех и более классов можно использовать тот же подход и попарно сравнивать классы друг с другом. Однако при этом может возникнуть проблема множественных сравнений. Поэтому для случая сравнения нескольких классов целесообразно использовать однофакторный дисперсионный анализ.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Дюк В.А., Самойленко А.П. Data mining: учебный курс.— СПб.: Питер, 2001.

2. Комлевая Н.О., Комлевой А.Н. Разработка информационной модели диагностирования состояния дыхательной системы // Холодильна техніка і технологія.— 2011.— Вып. 2(130).— С. 75 — 79.

3. Реброва О.Ю. Статистический анализ медицинских данных.— Москва: Медиа Сфера, 2006.

N.O. Komlevaya, A.Yu. Makhinenko, V.A. Lugovskoy, V.V. Provotorov

Using the Student's test for detecting statistically significant differences in classification problems.

The approach for using of statistical methods for objects classification is considered. Applying of Student's test allows comparing the mean values of attributes in classes. The statistical classifiers are formed for each class on the basis of the obtained results. The recommendations for deciding whether an object belongs to a certain class are given.

Keywords: *Student's test, classification, statistics.*