УДК 004:62-52:004.03

Н. О. КОМЛЕВАЯ[1], А. Н. КОМЛЕВОЙ[2], Б. И. ТИМЧЕНКО[3]

[1] ***Одесский национальный политехнический университет, Украина***
[2] ***Одесский национальный медицинский университет, Украина***
[3] ***Одесский национальный политехнический университет, Украина***

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ДВУХ ПОДХОДОВ ПРИ РЕШЕНИИ ЗАДАЧИ КЛАССИФИКАЦИИ

Проведен краткий обзор методов классификации. Детально проанализированы два подхода для решения задачи классификации: с использованием статистических методов и с помощью искусственных нейронных сетей. Для каждого из подходов рассмотрены математическая и программная реализации. Приведены практические результаты классификации с использованием специализированного приложения для статистического анализа STATISTICA и программной библиотеки для работы с искусственной нейросетью FANN (Fast Artificial Neural Network Library).

Ключевые слова: *статистические методы, искусственные нейронные сети, классификация, программные средства.*

Введение

Под классификацией понимают разделение множества объектов на подмножества по их сходству или различию в соответствии с принятыми методами. Классификация фиксирует закономерные связи между классами объектов. При этом под объектом понимается любой предмет, процесс, явление материального или нематериального свойства. Система классификации позволяет сгруппировать объекты и выделить определенные классы, которые будут характеризоваться рядом общих свойств. Таким образом, совокупность правил распределения объектов множества на подмножества считают системой классификации.

Для классификации используются различные методы [1], каждый из которых имеет свои преимущества и особенности применения. Основными из них являются 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 данной работы является рассмотрение двух подходов, которые можно использовать при решении задачи классификации: с использованием статистических методов и с помощью искусственных нейронных сетей. Проведенный анализ базируется на исследовании наборов медицинских данных, получаемых в результате специализированных обследований пациентов, которые входят в комплекс мер для проведения неинвазивного пульмонологического диагностирования. При этом полученные результаты используются для оценки состояния дыхательной системы обследованного пациента [2]. Выделение различных состояний дыхательной системы пациента в отдельные классы позволяет свести данную проблему к решению задачи классификации.

1. Описание принципа исследования изучаемых объектов

Диагностирование дыхательной системы пациента производится на основании анализа конденсата влаги выдыхаемого им воздуха; данный метод является достаточно новым, перспективным и неинвазивным [3].

В качестве материала для исследований взяты данные по трем группам пациентов, каждый из которых прошел пульмонологическое обследование. Промежуточные результаты обследования каждого пациента представлены вектором из 32 признаков, которые характеризуют состояние дыхательной системы. Значения признаков являются количественными величинами, измеренными на непрерывной шкале. Группировка пациентов проведена априорно на основании медицинских рекомендаций, основанных на стандартных диагностических методах; число пациентов в группах различно.

В данной работе предлагается выявить определенные закономерности в значениях признаков, что позволит считать сформированные группы не-

пересекающимися классами. Система классификации, основанная на этих закономерностях, позволит принимать решения о принадлежности изучаемых объектов тому или иному классу.

2. Статистический подход

Одним из методов многомерного статистического анализа является дискриминантный анализ, цель которого состоит в том, чтобы на основании значений признаков объекта классифицировать его некоторым оптимальным способом. Под оптимальным способом понимается либо минимум математического ожидания потерь [4], либо минимум вероятности ложной классификации.

Для каждой i-той рассматриваемой группы (i=1..3) имеется N_i пациентов (объектов исследования) с m=32 признаками. В результате измерений каждый объект характеризуется вектором $x_1 \ldots x_m$ значений признаков.

В первую очередь необходимо выполнить анализ количественных признаков трех выделенных групп путем проверки статистических гипотез. Это позволяет определить степень уверенности в том, что различия между генеральными совокупностями, из которых отобраны выборки признаков, действительно существуют. В качестве статистического метода выбран параметрический дисперсионный анализ по данным, полученным в несвязных (независимых) выборках.

Для расчета объёмов выборок на этапе планирования исследования определены следующие величины:

1) заданная мощность исследования (степень уверенности в том, что получен значимый результат, если он на самом деле имеет место в действительности);

2) уровень значимости σ – граничный уровень, ниже которого отбрасывается нулевая гипотеза;

3) величина изучаемого эффекта (насколько выражено различие между группами, которое нужно обнаружить и обосновать с помощью статистического анализа);

4) вариабельность изучаемых величин в группах.

Результаты измерения значения для каждого признака, согласно основной линейной математической модели дисперсионного анализа, получаются в результате сложения его точного значения, систематической и случайной ошибок измерения. В процедуре параметрического анализа вариаций общая вариация данных рассматривается как сумма двух видов вариаций – вариации между средним каждой группы и общим средним значением всей выборки (межгрупповая вариация), а также вариации между каждым объектом исследования группы и средним значением соответствующей группы (внутригрупповая вариация). Если межгрупповая вариация оказывается статистически значимо больше внутригрупповой вариации, то можно полагать, что различия между средними значениями групп существуют. Соотношение межгрупповой и внутригрупповой дисперсий имеет распределение Фишера и определяется при помощи F-критерия [5]. Расчеты были выполнены с использованием специализированного приложения для статистического анализа STATISTICA.

Таким образом, с помощью метода дисперсионного анализа были проверены нулевые гипотезы о том, что соответствующий признак не показывает различия между группами. В тех случаях, когда соответствующая нулевая гипотеза была отклонена, принималась альтернативная гипотеза о существовании различий между группами по исследуемому признаку. Возможность применения данного метода можно считать оправданной по следующим причинам:

– анализируемые признаки являются количественными,

– эти признаки нормально распределены в каждой из групп,

– дисперсии соответствующих анализируемых признаков в группах равны,

– группы определяются (детерминируются) качественным признаком.

Однако дисперсионный анализ позволил проверить лишь гипотезу об отсутствии различий между сравниваемыми группами по исследуемым признакам; с его помощью невозможно узнать, какие именно группы различаются между собой. Для выяснения этого были использованы методы множественных сравнений, являющихся частью, так называемого, апостериорного анализа (Post-hoc analysis). Механизм их работы заключается в проведении попарных сравнений средних значений всех групп, включенных в дисперсионный анализ.

Из набора тестов post-hoc анализа выбраны LSD (проверка критерия наименьшей значимой разности) и Newman-Keuls test (критерий Ньюмана-Кеулса). Согласно тесту LSD, на первом этапе метода при помощи критерия Фишера была проверена гипотеза о равенстве математических ожиданий распределений, из которых получены выборки значений признаков. Предварительно проверено, что эти выборки являются нормально распределенными и одноименные выборки обладают одинаковыми дисперсиями. Если гипотеза принималась, то метод останавливался, иначе переходили к следующему этапу. На втором этапе метода выборки были упорядочены по возрастанию выборочных средних и

далее поэтапно проверены гипотезы равенства средних соседних выборок при помощи критерия Стьюдента. В качестве оценки дисперсии использовалось внутригрупповое среднее. Если гипотеза о равенстве математических ожиданий принималась, то соответствующие выборки объединяются в одну группу.

Согласно методу Ньюмана-Кеулса, для выявления разницы между выборками рассчитывается значение стьюдентизированного диапазона на основании наибольших и наименьших средних значений выборок, дисперсий ошибок, которые берутся из таблицы ANOVA, и объемов выборок. Вычисленное значение сравнивается с критическим значением, взятым из таблицы распределения. Если вычисленное значение равно или больше, чем критическое значение, то нулевая гипотеза (о равенстве средних в выборках) может быть отклонена.

Для расчетов по обоим тестам использовались формулы, учитывающие неравные объемы выборок в разных группах. По результатам обработки данных о значениях признаков выяснилось, что существуют следующие категории признаков:

1) те, которые оказывают приблизительно сопоставимый вклад в объекты разных групп и, следовательно, не могут служить для различия этих групп (≈15%);

2) те, которые присутствуют только в одной группе и отсутствуют в других и в явном виде могут быть использованы для различия групп (≈20%);

3) те, которые присутствуют в разных группах, но оказывают различный вклад (≈65%).

Таким образом, предположение о различиях между выборками признаков в разных группах подтверждено, и группы можно считать непересекающимися классами.

Вернемся теперь к задаче классификации с использованием дискриминантного анализа, состоящей в том, чтобы по результатам измерений отнести объект к одному из классов. Используем тот факт, что число классов заранее известно, также известно, что объект заведомо принадлежит к определенному классу. Так как известно, что классы не пересекаются, требуется построить решающее нерандомизованное правило. Очевидно, при использовании решающего правила возникают потери, вызванные тем, что объект неправильно классифицирован – отнесен к одному классу, когда в действительности он принадлежит другому. Так как в нашем случае значение убытка при неправильной классификации объекта трудно оценить численно, то при построении оптимального правила был использован критерий минимальной вероятности ложной классификации.

Обработка данных была проведена с использованием двух методов дискриминантного анализа из пакета STATISTICA: стандартного и пошагового (включения и исключения). В качестве группирующей переменной выбирался диагноз пациента, в качестве независимых переменных – значения 32 признаков. Результаты дискриминантного анализа представляются в виде

$$class_i = \sum_{j=1}^{32} a_{ij} * x_j , \qquad (1)$$

где i – номер класса (i=1, 2, 3);

j – номер признака;

a_{ij} – коэффициенты решающего правила;

x_j – значение j-го признака для конкретного объекта.

Исследуемый объект относится к тому классу, для которого классификационное значение максимально. Кроме того, для классификации объекта можно использовать, так называемое, расстояние Махаланобиса, которое показывает квадрат расстояния от точки (объекта) до центров групп (классов). При этом объект относится к классу, до которого расстояние Махаланобиса минимально.

По результатам классификации 200 объектов, для которых был априорно известен $class_i$, с использованием описанного выше метода, в 198 случаях был получен истинный результат, для 2 случаев классификация была ложной.

3. Применение нейронных сетей

В задачах классификации и распознавания образов все чаще применяются искусственные нейронные сети (ИНС). Искусственной нейронной сетью называется математическая модель, программная, или аппаратная реализация, построенная по принципу функционирования и организации биологических нейронных сетей. Существует множество топологий нейронных сетей, как простейших (персептрон Розенблатта, сеть Ворда), так и сложных рекуррентных сетей. Среди различных структур нейронных сетей одной из наиболее известных является многослойная структура, в которой каждый нейрон произвольного слоя связан со всеми аксонами нейронов предыдущего слоя или, в случае первого слоя, со всеми входами ИНС. Нейронные сети не программируются в привычном смысле этого слова, они обучаются. Возможность обучения – одно из главных преимуществ нейронных сетей перед традиционными алгоритмами. Технически обучение заключается в нахождении коэффициентов связей между нейронами.

Для решения задачи классификации с использованием ИНС было использовано программное средство FANN (Fast Artificial Neural Network

Library) – бесплатная библиотека с открытым исходным кодом для создания нейронных сетей [6]. Сегодня библиотека FANN доступна почти для всех языков программирования и сред разработки. Она достаточно проста в использовании, универсальна и очень хорошо документирована.

Используемая библиотека включает в себя вариации градиентного алгоритма обучения персептрона методом обратного распространения ошибки. Основная идея этого метода состоит в распространении сигналов ошибки от выходов сети к её входам, в направлении, обратном прямому распространению сигналов в обычном режиме работы. Здесь используется метод обучения с учителем, поскольку известны как переменные стимулы (входные данные), так и переменные реакции, соответствующие этим стимулам (выходные данные). Библиотека предоставляет реализации следующих алгоритмов: инкрементальный – веса связей изменяются после каждого обучающего примера; групповой – веса связей изменяются после прохождения всех обучающих примеров; RPROP (Resilient PROPagation) – учитывается только знак производной от функции ошибки, находя веса итеративно при помощи последовательных приближений; QUICKPROP – используется метод касательных (метод Ньютона) для поиска минимума поверхности ошибок; SARPROP (Simulated Annealing Resilient PROPagation) – улучшение метода RPROP для избегания локальных минимумов поверхности ошибок.

Для классификации пациентов, прошедших пульмонологическое обследование, на основании значений их признаков были применены инкрементальный и групповой методы обратного распространения ошибки. Согласно их работе, после распространения входных сигналов через сеть вычисляется ошибка и передается обратно через сеть. В это время веса связей корректируются для уменьшения ошибки. Для рассматриваемых данных сеть имеет 32 входа, на которые подаются вещественные нормированные значения признаков, симметричные относительно нуля. ИНС имеет 4 выхода, из которых 3 соответствуют имеющимся классам, а последний соответствует ситуации, при которой исследуемый объект не принадлежит ни одному из описанных классов. Выходная информация содержит вероятности отнесения объекта к соответствующему классу. При большом количестве объектов, используемых на этапе обучения ИНС (>500), вероятность правильной классификации стремится к 1.

Алгоритм работы инкрементального и группового методов практически идентичен. Сначала входной сигнал распространяется через ИНС к выходным нейронам. Каждый скрытый нейрон вычисляет результат его активационной функции и рассылает свой сигнал всем выходным нейронам. После этого вычисляется ошибка e_k на одном выходном нейроне К:

$$e_k = d_k - y_k, \quad (2)$$

где d_k – желаемый, а y_k – действительный выход нейрона К.

После этого вычисляется δ_k, используемое для корректировки весов связей:

$$\delta_k = e_k g'(y_k), \quad (3)$$

где g' – производная от функции активации нейронов.

Согласно требованиям, функция активации в алгоритме обратного распространения ошибки должна обладать непрерывностью, дифференцируемостью и являться монотонно неубывающей, поэтому в качестве функции активации используется сигмоидальная функция:

$$f(x) = \frac{1}{1+e^{-x}}, \quad (4)$$

$$f'(x) = f(x)[1 - f(x)]. \quad (5)$$

Для каждого предыдущего слоя вычисляется δ_j на основе δ_k по формуле:

$$\delta_j = \eta g'(y_j) \sum_{k=0}^{K} \delta_k \omega_{jk}, \quad (6)$$

где К – количество нейронов в текущем слое;

ω_{jk} – веса входов;

η – параметр скорости обучения, который определяет, насколько сильно алгоритму следует изменять веса.

Этот процесс повторялся до тех пор, пока не было достигнуто некоторое условие установки, например достижение суммарной квадратичной ошибкой результата на выходе сети предустановленного заранее минимума в ходе процесса обучения, или выполнения определенного количества итераций алгоритма. По результатам классификации 200 объектов был получен результат, аналогичный предыдущему.

Заключение

Таким образом, в работе был проведен краткий обзор методов классификации, из всего разнообразия подходов подробно рассмотрены два: решение задачи классификации с использованием статистических методов и при помощи нейросетей. Оба подхода были использованы для проведения классификации состояния дыхательной системы пациентов.

Результаты показали высокую степень корреляции априорной и апостериорной классификацион-

ной информации. Использование в работе готовой свободно распространяемой библиотеки FANN позволит в дальнейшем разработать собственный программный продукт, предназначенный для проведения классификации на основании специализированных медицинских обследований.

Литература

1. Дюк, В. А. Data mining [Текст]: учебный курс / В. А. Дюк, А. П. Самойленко. – СПб. : Питер, 2001. – 412 с.

2. Комлевая, Н. О. Разработка информа-ционной модели диагностирования состояния дыхательной системы [Текст] / Н. О. Комлевая, А. Н. Комлевой // Холодильна техніка і технологія. – 2011. – Вып. 2(130). – С. 75 –79.

3. Комлевая, Н. О. Автоматизация диагностирования состояния дыхательной системы [Текст] / Н. О. Комлевая, А. Н. Комлевой // Труды тринадцатой МНПК «СИЭТ-2012». – Одесса, 2012. – С. 55.

4. Комлевая, Н. О. Построение системы диагностических признаков с использованием метода дискриминантного анализа в офтальмологических исследованиях [Текст] // Радіоелектронні і комп'ютерні системи. –2010. – № 6 (47). – С. 250 – 254.

5. Сепетлиев, Д. А. Статистические методы в научных медицинских исследованиях [Текст] / Д. А. Сепетлиев. – М. : Медицина, 1998. – 420 с.

6. Nissen, S. FANN – Fast Artificial Neural Network Library [Электронный ресурс] / S. Nissen. – Режим доступа: http://leenissen.dk/fann/wp/ – 5.03.2014.

Поступила в редакцию 05.03.2014, рассмотрена на редколлегии 25.03.2014

Рецензент: канд. техн. наук, доц., заведующий отделом информационно-телекоммуникационных технологий В. В. Капуа, Научно-производственная фирма Радуга-ТВ, Одесса, Украина.

ПОРІВНЯЛЬНИЙ АНАЛІЗ ДВОХ ПІДХОДІВ ПРИ ВИРІШЕННІ ЗАВДАННЯ КЛАСИФІКАЦІЇ

Н. О. Комлева, О. М. Комлевой, Б. І. Тимченко

Проведено короткий огляд методів класифікації. Детально проаналізовано два підходи для вирішення задачі класифікації: з використанням статистичних методів та за допомогою штучних нейронних мереж. Для кожного з підходів розглянуті математична і програмна реалізації. Наведено практичні результати класифікації з використанням спеціалізованої програми для статичного аналізу STATISTICA і програмної бібліотеки для роботи з штучною нейромережею FANN (Fast Artificial Neural Network Library).

Ключові слова: статистичні методи, штучні нейронні мережі, класифікація, програмні засоби.

COMPARATIVE ANALYSIS OF TWO APPROACHES IN SOLVING THE PROBLEM OF CLASSIFICATION

N. O. Komlevaya, A. N. Komlevoy, B. I. Tymchenko

The brief overview of the classification methods is done. Two approaches for solving the problem of classification are analyzed in detail: with using statistical methods and with using artificial neural networks. The mathematical and program implementations are given for each of these approaches. The practical results of the classifications are given with using the specialized software application for statistical analysis STATISTICA and the program library for working with artificial neural network FANN (Fast Artificial Neural Network Library).

Key words: statistical methods, artificial neural network, classification, software.

Комлевая Наталия Олеговна – канд. техн. наук, доцент, доцент кафедры системного программного обеспечения Одесского национального политехнического университета, Одесса, Украина, e-mail: nokoml@yandex.ua.

Комлевой Александр Николаевич – старший преподаватель кафедры клинической иммунологии, генетики и медицинской биологии Одесского национального медицинского университета, Одесса, Украина, e-mail: shurik-jan@yandex.ua.

Тимченко Борис Игоревич – Одесский национальный политехнический университет, Одесса, Украина, e-mail: tim4bor@gmail.com.