РАЗРАБОТКА ПРОГРАММНОГО ПРИЛОЖЕНИЯ ДЛЯ ОЦЕНКИ ИНФОРМАТИВНОСТИ ПРИЗНАКОВ (НА ПРИМЕРЕ ЗАБОЛЕВАНИЯ ЩИТОВИДНОЙ ЖЕЛЕЗЫ)

Прокопьев Р. О.
Научный руководитель: Берестнева О.Г. , д. т. н, профессор
Томский политехнический университет
tuz36@mail.ru

Введение

В наше время актуальной задачей является задача разработки приложений для оценки информативности признаков в задачах анализа экспериментальных данных для различных предметных областей.

В частности, оценка информативности признаков используется в медицине при диагностике многочисленных заболеваний. От результатов диагностики зависит дальнейшее лечение пациента. Поставить диагноз (то есть распознать то или иное заболевание или же его отсутствие) можно при условии, что проанализированы признаки, присущие объекту (в медицине – пациенту). Информативные показатели – это показатели, вносящие наибольший вклад в характеристику состояния объекта (пациента) [1].

Существуют многочисленные методы определения информативности признаков (метод кластеризации, метрика Кендалла – Кемен, статистический кластер – алгоритм и т.д.). Однако в отличие от других критериев статистической значимости различий, мера Кульбака позволяет оценить не достоверность различий между распределениями, а степень этих различий. Метод анализа признаков путем оценки информативности с использованием информативной меры Кульбака получил широкое применение в медицине при рассмотрении отдельных факторов, влияющих на постановку диагноза [2].

Описание метода

Метод Кульбака – предлагает в качестве оценки информативности меру расхождения между двумя классами, которая называется дивергенцией.

Согласно этому методу информативность или дивергенция Кульбака вычисляется по формуле [2]:

$$I(x_j) = \sum_{i=1}^{G} [P_{i1} - P_{i2}] \cdot \log_2 \frac{P_{i1}}{P_{i2}},$$

где G – число градаций признака;

P_{i1} – появление i-ой градации в первом классе.

$$P_{i1} = \frac{m_{i1}}{\sum_{i=1}^{G} m_{i1}},$$

где m_{i1} – частота появления i-ой градации в первом классе;

Знаменатель – появление всех градаций в первом классе, то есть общее число наблюдений в первом классе.

P_{i2} – появление i-ой градации во втором классе.

$$P_{i2} = \frac{m_{i2}}{\sum_{i=1}^{G} m_{i2}},$$

где m_{i2} – частота появления i-ой градации во втором классе;

Знаменатель – появление всех градаций во втором классе, то есть общее число наблюдений во втором классе.

Разработка программного продукта

Опираясь на описанный ранее метод Кульбака, мы разработали специализированную программу для оценки информативности признаков «Informative features».

Для создания программного приложения использовалась среда C++ Builder 6. Программа создана в виде оконного приложения. Рабочая область программы представлена на рисунке 1.

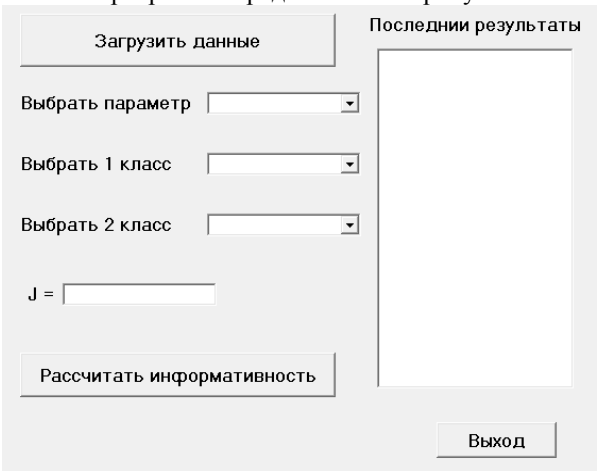

Рис. 1. Рабочая область программы «Informative features»

Для того, чтобы начать работать с программой, необходимо нажать на кнопку «Загрузить данные». При нажатии на эту кнопку, пользователь загружает имеющийся на компьютере Excel файл, который включает в себя список пациентов с соответствующими им параметрами (признаками) и принадлежностью к какому – либо классу, между которыми и рассчитывается информативность. Далее для работы с программой необходимо выбрать параметр, для которого необходимо рассчитать информативность. Завершающим этапом является

выбор пользователем двух классов из всего имеющегося набора классов (диагнозов), после чего пользователь и нажимает на кнопку «Рассчитать информативность».

Программа производит расчет и выдает результат. В окне справа разработанный продукт «Informative features» отображает все произведенные вычисления. При этом результат каждого нового расчета отображается в окне справа, таким образом, пользователь может сравнивать информативность каждого признака.

Апробация программного продукта была произведена в НИИ курортологии и физиотерапии г. Томска (на клинических данных пациентов с ожирением и заболеваниями щитовидной железы).

Щитовидная железа – один из важнейших органов внутренней секреции человека. Особенно велико ее значение для развивающегося, растущего организма. Физиологическое действие тироидных гормонов разнообразно и направлено на все обменные процессы, функции многих органов и тканей, в том числе на развитие плода, процессы роста и дифференцировки тканей, особенно нервной системы. В отличие от взрослых, тироидная недостаточность у детей раннего возраста резко задерживает рост скелета и созревание ЦНС. Только раннее и адекватное лечение пациентов с подобным заболеванием тироидными препаратами обеспечивает хороший прогноз физического и умственного развития у детей с врожденным гипотиреозом [3].

Примеры работы программы для оценки информативности показателей ОТ (объем талии) и ИМТ (индекс массы тела) приведены на рисунках 2-3.

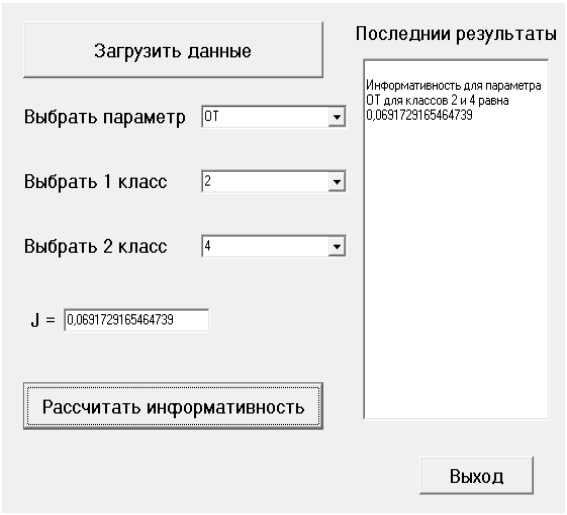

Рис. 2. Пример работы программы «Informative features»

На рисунке 2 приведен пример расчета информативности для параметра «объем талии» между классами 2 и 4.

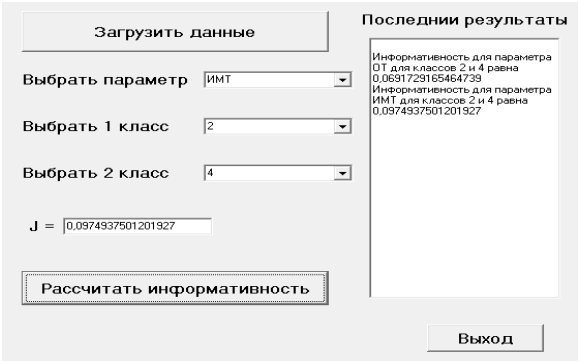

Рис. 3. Пример работы программы «Informative features»

На рисунке 3 приведен пример расчета информативности для параметра «индекс массы тела» между классами 2 и 4.

Из полученных результатов можно сделать вывод, что между выбранными параметрами с принадлежностями 2 и 4 классу, «индекс массы тела» имеет информативность больше, чем параметр «объем талии».

Заключение

Оценка значимости того или иного признака дает возможность сократить число ошибок при решении любой классификационной задачи, так как позволяет отобрать менее информативные признаки от более информативных. В частности признак может иметь большую диагностическую ценность в сравнении с каким-либо другим.

Разработанное приложение оценивает информативность признаков при дифференциальной диагностике заболеваний.

Приложение апробировано и внедрено в опытную эксплуатацию в НИИ курортологии и физиотерапии г Томска (для определения наиболее информативных признаков при диагностике заболеваний щитовидной железы. Программа может быть адаптирована для решения задач диагностики других заболевания или же для оценки информативности признаков в других сферах деятельности.

Исследование выполнено при финансовой поддержке РФФИ, в рамках научного проекта 15-07-08922

Список использованных источников

1. Берестнева О.Г., Шаропин К.А., Старикова А.В., Кабанова Л.И. Технология формирования баз знаний в медицинских информационных системах//Известия Южного федерального университета. Технические науки. – 2010. Т.109, № 8. – С. 32-37.
2. Гублер Е. В. Вычислительные методы анализа и распознавания патологических процессов. – М.:МЕДИЦИНА, 1978. – 198с.
3. Русский Медицинский Сервер – ТИРОНЕТ – все о щитовидной железе. [Электронный ресурс]. – Режим доступа: http://thyronet.rusmedserv.com/, свободный (дата обращения: 4.08.2015).