ОЦЕНКА ИНФОРМАТИВНОСТИ МНОГОМЕРНЫХ ДАННЫХ

Емельянова Ю.А.
Научный руководитель: Марухина О.В., доцент каф.ПИ
Томский политехнический университет, 634050, Россия, г. Томск, пр. Ленина, 30
E-mail: yuliyaemelianova@yandex.ru

Введение

В настоящее время вся информация хранится в электронном виде в базах данных и занимает большие объемы. Любой объект характеризуется некоторым числом параметров. Многомерные данные содержат информацию о трех или более признаках для каждого объекта. Эти данные в дальнейшем могут использоваться для получения информации о зависимостях между признаками.

Проблема обработки многомерных данных

Данные могут содержать большое количество скрытых закономерностей, которые являются важными для принятия стратегических решений. Следовательно, существует необходимость анализировать подобного рода данные и представлять новые знания в удобной для восприятия человеком форме, хранить только те данные, которые несут в себе необходимую информацию, возможно, некоторые из них могут просто занимать место на компьютере и не нести в себе смысла.

За последнее время анализ многомерных данных стал активно развивающимся и применяющимся практически во всех областях исследований. Анализ многомерных данных является одной из наиболее востребованных междисциплинарных областей знания.

Любая обработка информации в различных областях, таких как медицина, банковское дело, телекоммуникации, молекулярная генетика посвящена конкретным целям, например, в медицине – это исследование болезней, лечение, анализ, постановка диагноза у больного. Поэтому признаки, присущие пациенту, которые получены и проанализированы, являются информативными признаками, с помощью которых можно распознать заболевание или его отсутствие у пациента. Важной задачей является поиск и отбор признаков достаточно информативных для постановки достоверного диагноза.

Описание объектов содержат все доступные наблюдению, измерению параметры, характеристики, признаки, поэтому в описании используется множество величин. Такой большой набор данных требует трудоемких работ при обработке данных. При принятии решения о выборе класса, которому принадлежит анализируемый объект, возникает проблема его оценки по нескольким признакам. Это проблема делится на подпроблемы: установление обобщенного признака и определение важности признаков, отражающих свойства объектов.

Существуют параметрические и непараметрические методы оценки информативности. Параметрические методы основаны на предположениях о характере распределения случайной величины и делаются предположения как параметры в разных выборках, соотносятся между собой. Непараметрические методы математической статистики – методы, при которых не выдвигаются какие-либо предположения о характере распределения исследуемых данных.

Наиболее известные методы оценки информативности. Метод накопленных частот – при таком методе берутся две выборки признака, принадлежащие двум различным классам, по обеим выборкам в одних координатных осях строят эмпирические распределения признака и подсчитывают накопленные частоты (сумму частот от начального до текущего интервала распределения). Оценкой информативности служит модуль максимальной разности накопленных частот. В методе Шеннона информативность оценивается как средневзвешенное количество информации, приходящееся на различные градации признака.

Оценка информативности по Кульбаку

Метод Кульбака предлагает в качестве оценки информативности – меру расхождения между двумя классами. Согласно этому методу информативность вычисляется по формуле 1:

$$J(x_i) = \sum_i 10\lg\frac{P(x_{ij}/A_1)}{P(x_{ij}/A_2)}0{,}5[P\left(\frac{x_{ij}}{A_1}\right) - P(x_{ij}/A_2)], \quad (1)$$

$J(x_i)$ –информативность признака,

P_1- вероятность попадания признака в первом классе A_1,

P_2- вероятность попадания признака во втором классе A_2,

j – номер диапазона признака x_i .

$J(x_i)$ – величина всегда будет положительной. Это связано со свойством логарифмов. Если числитель логарифмической дроби больше знаменателя, то логарифм отношения вероятностей будет положительной величиной. Если числитель логарифмической дроби меньше знаменателя, то логарифм отношения вероятностей и разность вероятностей будут отрицательными величинами и при перемножении дадут положительную величину. Таким образов, величина $J(x_i)$, будучи всегда положительной, отразит абсолютное значение вклада данного диапазона в приближении к любому правильному диагностическому порогу [1].

Были предоставлены данные по ожирению. Данные имеют вид таблиц с пациентами и

множеством признаков. Определение информативной ценности связано с необходимостью выделения наиболее информативных признаков, получаемых при обследовании больных. Необходимо было рассчитать информативность для следующих групп: клиника, сердечно-сосудистая система, физическая работоспособность, липидный обмен, биохимия крови, гормональный статус, иммунологический статус, состояние калликреин-кининовой системы и окислительная способность плазмы крови. Каждая группа имеет списки признаков. Этапы расчёта информативности:

1. Были взяты признаки до и после лечения из одной группы.
2. Далее определили для каждого признака минимальное и максимальное значение.
3. Задали количество интервалов распределения и подсчитали количество значений, которые попадают в каждый интервал (всего 5 интервалов распределения).
4. Рассчитали вероятность попадания признака в группу 1 (до лечения) и в группу 2 (после лечения).
5. По формуле, приведенной выше, рассчитали информативность.

Информативность групп были получены в результате выполнения программы, созданной командой, занимающаяся данной темой в ТПУ. В результате была посчитана информативность для каждой группы признаков. Например, группа «Физическая работоспособность» включает в себя следующие признаки: толерантность к физической нагрузке, общая работоспособность, индекс инсулинорезистентности и двойное произведение (насыщение миокарда кислородом). Информативность данной группы представлена в таблице 1.

Таблица 1. Информативность группы

Признак	**Информативность**
ТФН	1,4
Общая раб-ть	0,99
HOMA	0,44
Дв.Пр.	0,01

Из таблицы видно, что информативным признаком в данной группе является ТФН (рис.1), менее информативным – Дв.Пр.(рис.2).

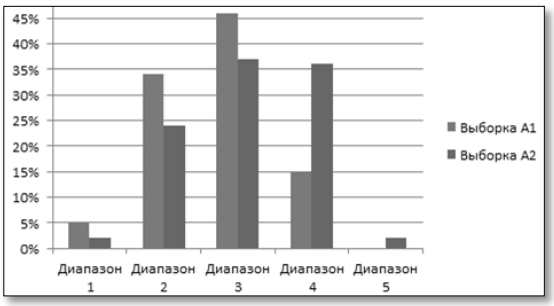

Рис.1. Признак «ТФН»

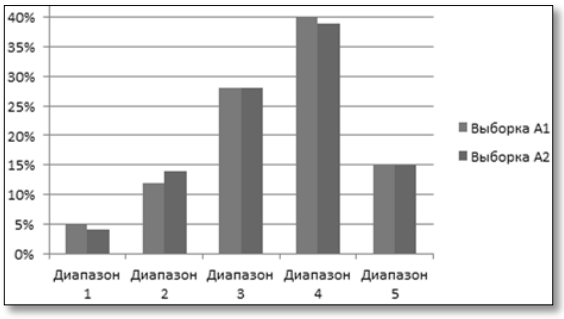

Рис.2. Признак «Дв.Пр»

Для каждой группы были построены графики наиболее информативного и менее информативного признака. Выборка А1 – значения до лечения, выборка А2 – значения после лечения. Из графиков видно, что чем больше разница между выборками, тем информативнее признак.

Заключение

В результате рассчитана информативность для различных групп данных по ожирению. Получили, что наиболее информативными являются следующие признаки: систолическое артериальное давление, толерантность к физической нагрузке, липопротеиды низкой плотности, щелочная фосфатаза в сыворотке крови, циркулирующие иммунные комплексы, уровень каллекриина и содержание оксида азота в крови. Метод Кульбака служат для определения информативности признака, который участвует в распознавании двух классов объектов.

Список литературы:

1. Е.В.Гублер. Вычислительные методы анализа и распознавания патологических процессов, 1978, 269 с.
2. И.С. Голованова. Выбор информативных признаков. Оценка информативности. [Электронный ресурс]. URL: http://ime.tpu.ru/study/discypliny/INF-PR.pdf (дата обращения: 11.05.2017).
3. Капустина С.В., Кирякова О.В., Лапина Л.А. Выбор информативных признаков для тяжести заболевания. [Электронный ресурс]. URL: https://www.science-education.ru/ru/article/view?id=21955 (дата обращения: 17.05.2017).
4. Колесникова С.И. Методы анализа информативности разнотипных признаков. [Электронный ресурс]. URL: http://cyberleninka.ru/article/n/metody-analiza-informativnosti-raznotipnyh-priznakov (дата обращения: 17.05.2017).
5. Оценка информативности признаков в генетике. Оценка биомедицинских измерений. [Электронный ресурс]. URL: http://dommedika.com/235.html (дата обращения: 11.05.2017).