Учеными Лундского университета было выполнено исследование [1] с целью создания более точной и эффективной классификации типов сахарного диабета, после публикации которого возникли основания говорить о большем количестве типов сахарного диабета. Авторы научной работы [1] выявили, что в структуре анализируемых данных о W (W = 8980) пациентах с сахарным диабетом из реестра «The Swedish All New Diabetics in Scania» (ANDIS) хорошо различимы 5 кластеров. Этот факт указывает на необходимость выделения 5 типов сахарного диабета вместо классификации с 2 основными и несколькими дополнительными типами. Однако, авторами статьи [1] не было учтено влияние случайных погрешностей исходных медицинских данных, в частности инструментальных погрешностей, на целесообразность выделения итогового количества кластеров. Таким образом, возникает риск, что полученная кластерная структура есть результат причудливой реализации случайных погрешностей. Данное обстоятельство требует тщательного изучения, что и было произведено авторами этой работы.

В работе [1] каждому i-ому пациенту (i = 1, 2, …, W) из реестра ANDIS был сопоставлен вектор **z**i = (HbA1c, BMI, age, HOMA2-B, HOMA2-IR, GADA) числовых характеристик его физиологического состояния: первые 5 параметров – количественные (параметры HOMA2-B и HOMA2-IR вычислялись на основании значений концентраций глюкозы (glucose) и С-пептида (C-peptide) с помощью HOMA2 calculator (University of Oxford, UK)), последний (GADA) – бинарный. После чего, получившийся набор векторов Z = (**z**1, **z**2, …, **z**W)T был подвернут кластеризации методами k-means и TwoStep.

Для исследования влияния случайных погрешностей исходных данных на степень обоснованности принимаемой классификации типов сахарного диабета необходимо было оценить наибольшее число значимо отличающихся кластеров (numcl) при кластеризации набора векторов числовых характеристик, описанных ранее, с учетом случайных погрешностей, методами k-means и TwoStep.

В данной работе при кластеризации методом k-means оценка numcl была выполнена согласно алгоритму [2], а при кластеризации методом TwoStep значение numcl соответствовало числу кластеров с наименьшим значением Байесовского информационного критерия (BIC).

Автором данной работы было рассмотрено влияние только инструментальных погрешностей на итоговую кластерную структуру. В статье [1] не были упомянуты характеристики инструментальных погрешностей, поэтому автором этой работы было направлено соответствующее письмо Лейфу Групу, одному из авторов работы [1], на который получен любезный ответ, содержащий необходимую информацию о характеристиках инструментальных погрешностей. Основной характеристикой инструментальных погрешностей в работе [1] являлся коэффициент вариации (cv): = 1 % (минимальное значение cv среди используемых средств измерений); = 4 %; = 3 %.

Оценка numcl сводится у следующей последовательности действий:

1. *Оценка характеристик погрешностей параметров HOMA2-B и HOMA2-IR.*

Параметры HOMA2-B и HOMA2-IR являются неточными, так как вычисляются на основании параметров глюкоза и С-пептид, которые содержат инструментальную погрешность. На основании значений характеристик погрешностей параметров глюкоза и С-пептид с помощью метода Монте-Карло, где количество повторов N = 103, были оценены погрешности параметров HOMA2-B и HOMA2-IR. Оценка была выполнена путем вычисления второго начального момента погрешности: , .

1. *Постановка математического моделирования.*

Был сгенерирован набор векторов модельных значений количественных параметров Z: , где с – координаты центра кластера («Table S2. Patient characteristics in ANDIS using k-means clustering» [1]), σ – среднеквадратическое отклонение, которое вычислялось следующим образом: (l и r – границы диапазонов количественных параметров («Figure 2: Cluster characteristics in the ANDIS cohort» [1])). . Бинарный параметр GADA был установлен положительным, аналогично данным из работы [1].

1. *Моделирование инструментальных погрешностей.*

Было произведено моделирование инструментальных погрешностей количественных параметров. В результате были получены N наборов инструментальных погрешностей (j = 1, 2, …, N), где . Моделирование производилось согласно следующим условиям:; в работе [1] , что соответствует и , (MV – максимальная вариация веса человека за день), , ; ; ; .

На основе полученных наборов инструментальных погрешностей количественных параметров и набора модельных данных были вычислены наборы неточных данных .

1. *Оценка numcl.*

Наборы неточных данных количественных параметров были кластеризованы методами k-means и TwoStep. Оценка numcl была произведена согласно описанию, представленному выше.

Реализация данного алгоритма была выполнена на языке программирования Python и в системе математического моделирования Matlab

Результаты (метод k-means):

|  |  |  |
| --- | --- | --- |
| **Число кластеров** | **min** | **max** |
| 1 | 15987359 | 16101733 |
| 2 | 7933740 | 7990532 |
| 3 | 4176331 | 4234761 |
| 4 | 2769665 | 2803067 |
| ***5*** | 2369259 | 2422290 |
| 6 | 2002436 | 2040794 |
| 7 | 1669489 | 1697011 |
| 8 | 1479866 | 1520731 |
| 9 | 1316727 | 1405255 |
| 10 | 1208059 | 1268550 |
| 11 | 1125257 | 1185244 |
| 12 | 1056133 | 1102497 |
| 13 | 996995 | 1049026 |
| 14 | 945928 | 994167 |
| 15 | 895812 | 944402 |
| **16** | ***853811*** | 892871 |
| 17 | 820346 | ***857331*** |

Исходя из получившихся результатов, можно сделать вывод о корректности итоговой кластерной структуры в исследовании [1], так как с учетом инструментальных погрешностей, получившееся число кластеров, равное 16, значительно превышает число кластеров из работы [1], равно 5.

1. Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A. et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. The lancet Diabetes & endocrinology. 2018. Vol. 6. No 5. P. 361-369. URL: <https://www.thelancet.com/journals/landia/article/PIIS2213-8587(18)30051-2/fulltext>
2. Semenov, K. K., Bolschikov, V. A. (2019). The metrologically reasonable clustering of measurement results. Journal of Physics: Conference Se-ries. Vol. 1379, No. 1, p. 012054.