Практические задания по дисциплине «Основы интеллектуального анализа данных» (2022)

Схема эксперимента на модельных данных

Упрощенная онтология медицинской диагностики

База знаний

Неинтересные параметры	Интересные параметры
• классы (заболевания); • признаки; • возможные значения признаков; • клинические картины заболеваний.	• нормальные значения признаков; • число периодов динамики признака при заболевании; • значения для каждого периода признака при заболевании; • верхняя граница длительность каждого периода признака при заболевании; • нижняя граница длительность каждого периода признака при заболевании.

Выборка данных (истории болезни)

Наблюдаемые значения	Ненаблюдаемые значения
• диагноз; • наблюдаемые признаки; • моменты наблюдения признаков (с момента начала заболевания); • значения признаков в моменты наблюдения.	• разбиение каждого наблюдаемого признака на периоды динамики.

Задание 1. Генерация модельной базы знаний (МБЗ)

1.1. Сгенерировать в Excel модельную базу знаний на основе упрощенной онтологии медицинской диагностики.

1.2. Основные условия генерации:
- количество классов (заболеваний) = 2;
- количество признаков = 6 (по два каждого типа: бинарный, перечислимый, интервальный);
- число периодов динамики (ЧПД) для каждого признака генерируется из интервала [1, 5];
- ЧПД должно быть равно 1 в редких случаях (не более одного);
- количество одноименных признаков, у которых совпадают ЧПД в разных заболеваниях = 3;
- значения в соседних периодах признака – не пересекаются;
- нижняя граница периода не превышает его верхнюю границу;
- нижняя граница равна 1 в редких случаях (не более одного).

1.3. Приложить файл *.xls (xlsx) к заданию в Teams и отметить задание как сданное – оно будет оценено бинарно (1 – сдано, 0 – нет).

Задание 2. Генерация модельной выборки данных (МВД)

2.1. На основе сгенерированной вами в задании 1 модельной базы знаний, сгенерировать модельную выборку данных.

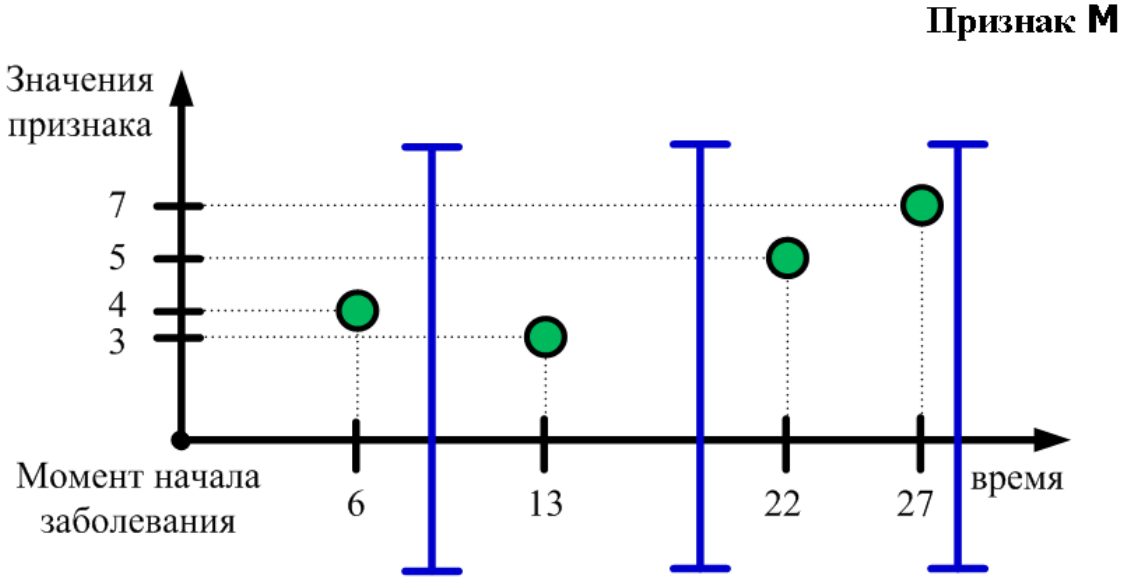
2.2. Основные условия генерации:
- длительность периода динамики должна находиться в интервале [нижняя граница, верхняя граница];
- количество моментов наблюдения в периоде динамики – от 1 до 3;
- количество историй болезни (ИБ) – по 3 на каждое заболевание.

2.3. Приложить файл *.xls (xlsx) к заданию в Teams и отметить задание как сданное – оно будет оценено бинарно (1 – сдано, 0 – нет).

Задание 3. Формирование альтернатив индуктивной базы знаний (ИФБЗ)

На основе сгенерированной вами в задании 2 модельной выборки данных, индуктивно сформировать альтернативы индуктивной базы знаний.

3.1. Расставить границы периодов динамики для 2-х любых признаков 2-х любых историй болезни по каждому заболеванию. Условия расстановки границ периодов динамики для признака из одной истории болезни:
- варианты расстановок границ сформировать для всех ЧПД, принадлежащих [1, 5];
- в каждый период динамики должен попасть минимум 1 момент наблюдения;
- граница между двумя моментами наблюдения ставится посередине;
- граница после последнего момента наблюдения ставится в следующий за ним час.



Число периодов динамики = 3

3.2. На основе каждой расстановки границ периодов динамики сформировать альтернативу индуктивной базы знаний (эта альтернатива будет относиться к конкретному признаку из конкретной истории болезни). Условия формирования:

- для каждого периода динамики формируется значение параметра «Значения для периода», в него включаются все неповторяющиеся значения в моменты наблюдения, которые попали в этот период;
- для каждого периода динамики формируется значение параметра «Нижняя граница» как разница первого момента наблюдения в периоде и левой границы периода;
- для каждого периода динамики формируется значение параметра «Верхняя граница» как разница последнего момента наблюдения в периоде и левой границы периода.

3.3. Объединить альтернативы по одноимённым признакам у двух историй болезни с одним диагнозом. Условия:
- объединяются альтернативы с одинаковым ЧПД;
- при объединении для соответствующих периодов «Значения для периода» объединяются;
- при объединении для соответствующих периодов «Нижняя граница» выбирается минимальной из двух;
- при объединении для соответствующих периодов «Верхняя граница» выбирается максимальной из двух;
- после объединения сохраняется только результат объединения (объединявшиеся альтернативы удаляются).

3.4. Сократить множество объединённых альтернатив: если в соседних периодах динамики «Значения для периода» пересекаются, удалить эту альтернативу (пометить красным цветом фона).

3.5. Выделить (цветом и комментарием рядом) альтернативу для каждого признака, которая больше всего похожа (с точки зрения ЧПД и значений для периода) на описание этого признака в МБЗ.

3.6. Приложить файл *.xls (xlsx) к заданию в Teams и отметить задание как сданное – оно будет оценено бинарно (1 – сдано, 0 – нет).

Задание 4. Сравнение МБЗ и ИФБЗ

Сравнить сгенерированную в задании 1 модельную базу знаний (МБЗ) и, сформированную в задании 3, индуктивную базу знаний (ИФБЗ).

Результаты сравнения:

4.1. Процент совпадения ЧПД у одноименных признаков – отдельно для каждого заболевания.

Пример:

МБЗ

Заболевания	*Признаки*	*ЧПД*
Заболевание 1	*Признак 1*	*3*
Заболевание 1	*Признак 2*	*3*
Заболевание 1	*Признак 3*	*4*
Заболевание 2	*Признак 1*	*1*
Заболевание 2	*Признак 2*	*5*
Заболевание 2	*Признак 3*	*3*

ИФБЗ

Заболевания	*Признаки*	*ЧПД*
Заболевание 1	*Признак 1*	*3*
Заболевание 1	*Признак 2*	*3*
Заболевание 1	*Признак 3*	*4*
Заболевание 2	*Признак 1*	*2*
Заболевание 2	*Признак 2*	*5*
Заболевание 2	*Признак 3*	*3*

Процент совпадения ЧПД:

- *заболевание 1 – 100%*
- *заболевание 2 – 66,6%*

4.2. Средний процент совпадения ЧПД у одноименных признаков – для всех заболеваний.

Пример: предыдущий.

Средний процент совпадения ЧПД:

- *(100+66,6)/2=83,3%*

4.3. Соотнесение областей значений признаков (ЗДП) в соответствующих периодах (только для одноименных признаков, у которых ЧПД совпали в МБЗ и ИФБЗ) – отдельно для каждого заболевания. Считается по каждому периоду динамики.

Пример:

МБЗ

Заболевания	*Признаки*	*ЧПД*	*ПД*	*ЗДП*
Заболевание 1	*Признак 1*	*3*	*1*	*{а, б, в, г}*
			2	*{д, е, ж}*
			3	*{в, з}*
Заболевание 1	*Признак 2*	*3*	*1*	*0-10*
			2	*15-20*
			3	*0-3*

ИФБЗ

Заболевания	*Признаки*	*ЧПД*	*ПД*	*ЗДП*
Заболевание 1	*Признак 1*	*3*	*1*	*{а, б, в, г}*
			2	*{д}*
			3	*{в, з}*
Заболевание 1	*Признак 2*	*3*	*1*	*3-7*
			2	*11-20*
			3	*33-35*

Процент тождественного совпадения $ЗДП_{МБЗ}$ и $ЗДП_{ИФБЗ}$ (зеленые):

- *заболевание 1: (2/6)=33,3%*
  (для признака 1 полностью совпали ЗДП в периодах: 1 и 3,
  для признака 2 полного совпадения нет, всего периодов – 6).
- *заболевание 2: …*

Процент $ЗДП_{ИФБЗ}$ подмножество $ЗДП_{МБЗ}$ (синие):

- *заболевание 1: (2/6)=33,3%*

Процент $ЗДП_{МБЗ}$ подмножество $ЗДП_{ИФБЗ}$ (желтые):

- *заболевание 1: (1/6)=16,6%*

Все оставшиеся случаи (красные):

- *заболевание 1: (1/5)=16,6%*

4.4. Средний процент соотнесения областей значений признаков (ЗДП) в соответствующих периодах (только для одноименных признаков, у которых ЧПД совпали в МБЗ и ИФБЗ) – для всех заболеваний.

4.5. Приложить файл *.xls (xlsx) к заданию в Teams и отметить задание как сданное – оно будет оценено бинарно (1 – сдано, 0 – нет).