**Формирование и анализ наборов информативных признаков объектов по парам классов**

Н.А ИгнатьевI, М.А. РахимоваI

IНациональный Университет Узбекистана им. Мирзо Улугбека, г. Ташкент, Узбекистан

**Аннотация.** Рассматривается формирование информативных наборов признаков, отбор которых производится по парам непересекающихся классов объектов. Считается, что индексы классов принадлежат множеству допустимых значений классификационного признака в порядковой шкале. Процедура отбора реализуется алгоритмом иерархической агломеративной группировки путем разбиения множества исходных признаков на непересекающиеся группы (наборы признаков). Для снижения размерности пространства в описании объектов используется вычисление латентных показателей по группам. Определена мера компактности классов по латентному показателю для оценки набора признаков. Оценка набора с максимальным значением меры компактности является условием выбора его в качестве информативного. Исследуется свойства бинарных отношений между классами по значениям меры компактности.

**Ключевые слова**: нелинейные преобразования, функции принадлежности, иерархическая агломеративная группировка, обобщённые оценки объектов, мера компактности.

**Введение**

Для обоснования принимаемых решений в слабо формализуемых предметных областях часто используют упорядочение объектов и разделение их на классы. Наличие информативных наборов признаков имеет существенное влияние на повышение обоснованности решений. Использование единого набора признаков не всегда является эффективным. При числе классов больше 2 поиск единого набора рассматривается как попытка найти “*усреднённое”* решение без учёта специфики классов. Проблемы отбора возникают и при большой размерности (20 тыс. и более) пространства.

Упорядочение объектов является распространённым явлением на практике. В медицине с целью выбора технологии лечения применяют ранжирование пациентов по степени тяжести заболевания, учебные группы в образовательных учреждениях формируют по оценкам студентов при сдаче минимума на знание иностранного языка. При идентификации (вербальном описании) классов их упорядоченность облегчает процесс интерпретации пользователем результатов анализа данных на компьютере. Примером может служить исследование описаний 4 групп больных с бессимптомной, лёгкой, среднетяжёлой и тяжёлой формой COVID–19. Потребность в использовании нескольких наборов информативных признаков необходима для объяснения роста (количественной оценки) тяжести заболевания при сравнении описаний пациентов от бессимптомной до тяжёлой формы. В качестве данных для вычисления количественных оценок можно использовать значения латентных показателей, формируемых по информативным наборам признаков.

Для сравнения, упорядочения и классификации объектов по их свойствам в [1] предложена мультиметодная технология многокритериального выбора. Критерии использовались для последовательного снижения размерности признакового пространства путем агрегирования исходных характеристик объектов (числовых, символьных или вербальных). Шкалы критериев формировались с помощью разных комбинаций методов вербального анализа решений на основе знаний экспертов и/или предпочтений лица принимающего решение (ЛПР).

Общепринятой методики для анализа разнотипных (количественных и качественных) данных не существует. При объяснении результатов анализа количественных (сырых и латентных) данных часто пользуются отношением «показатель больше (меньше) нормы». Как правило, указывается интервал, в границах которого находится значения нормы. В медицинской практике существует альтернативная точка зрения, что у каждого человека есть своя «норма здоровья» [2], которая не обязательно совпадает с официально принятыми показателями для оценки состояния здоровья. Для обоснования этой точки зрения имеет смысл проверка гипотезы о существовании нескольких интервалов, в границах которых находятся значения нормы для отдельных групп людей. Например, представителей мировой элиты по некоторым видам спорта. Есть потребность в разработке и использовании специальных критериев для вычисления границ интервалов.

При выборе нового пространства для описания объектов в двухклассовой задаче распознавания предложено два критерия для разбиения значений признаков на непересекающиеся интервалы [3]. Два свойства определяют различие между этими критериями: порядок следования значений признака; частота встречаемости значений признака в описании объектов из каждого класса. По первому свойству определяется мера различия (компактность) признака и граница (порог) между классами. Второе свойство используется для вычисления значений функции принадлежности объектов к классам [4].

Замена исходных значений признака на значение функции принадлежности объектов к классам рассматривается как нелинейное преобразование. Для номинальных признаков это возможность использовать вместо градаций описание объектов в порядковой шкале, либо уменьшить число градаций до двух. Описание процесса перехода от значений исходного “*сырого*” признака до представления его через две градации определяются как действия по формированию ансамбля элементарных классификаторов для построения метамодели алгоритмов распознавания с применением технологии стекинга [2, 5].

Предлагается процесс отбора информативных наборов разнотипных признаков реализовать через синтез латентных показателей (признаков) по парам классов. Рассматривается специальная мера для определения различия между объектами классов по латентным признакам. Ожидаемые результаты отбора информативных признаков при числе классов больше 2:

– набор сырых признаков общий для всех классов;

– наборы сырых признаков разные по каждой паре классов.

Набор сырых признаков общий для всех классов рассматривался в [6]. Значения латентных показателей по набору из качественных и количественных синдромов и симптомов на данных 14 возрастных групп [7] интерпретировались как количество здоровья. Экспериментально было подтверждено, что с увеличением возраста количество здоровья по возрастным группам монотонно уменьшается.

Использование латентных признаков в данной работе связано с вычислением их значений как обобщённых оценок объектов по парам классам. Предполагается, что свойства объектов непересекающихся классов определяют оппозицию друг к другу (больной – здоровый; сильный – слабый; бедный – богатый). Вычисление обобщённых оценок по наборам сырых признаков может производится как с использованием их предобработки так и без неё. При предобработке значения сырых признаков в описании объектов отображаются в {1,2} с сохранением размерности пространства.

Перспективы использования латентных признаков на базе обобщённых оценок объектов связаны с разработкой критериев для группировки (формирования наборов) исходных признаков с целью сокращения размерности пространства в двухклассовой задаче. По результатам применения критериев формируется непустое множество непересекающихся наборов признаков. Одной из проблем для исследования является доказательство существования набора признаков, при вычислении обобщённых оценок по которому точность разделения объектов на два класса выше чем по обобщённым оценкам, вычисляемым по всему множеству признаков. Доказательство существования актуально и для данных с большой размерностью пространства, решение которых известными методами невозможно.

1. **Постановка задачи**

Дана выборка объектов *E*0={*S*1,…,*Sm*}, разделённая на *l*(*l*>2) непересекающихся классов *K*1,…,*Kl*. Считается, что значения индексов классов в {1,…,*l*} принадлежит множеству допустимых преобразований классификационного признака в порядковой шкале. Описание объектов в *E*0 представлены набором разнотипных признаков *X*(*n*)=(*x*1,…,*xn*), σ из которых являются номинальными, *n*-σ – количественными.

Определены процедуры для:

– разбиения *X*(*n*) на непересекающиеся подмножества *Xij*(*n*1),…, *Xij*(*nt*) по паре классов (*Ki*,*Kj*), *i*<*j*, *t*≥1, Ɐ*c*∈{1,…,*t*} *nc* >1, *n*1+…+*nt* ≤*n* и вычисления по ним наборов латентных признаков ;

– вычисления значений меры компактности , *c*∈{1,…,*t*} на *Ki*U*Kj* по  и 

Требуется:

–для каждого *i*∈{1,…,*l* -1} произвести отбор множества



–сформировать матрицу Α={α*ij*}*i,j*∈{1,…,*l*}, где 

–выделить Ω*i*, *i*∈{1,…,*l-*2}, для которых α*i,i*+1 *H* α*i,i*+2 = …= α*i,l*-1 *H* α*il* =true, где *H*∈{≤, ≥}.

**2. Разбиение значений признаков на интервалы**

Для анализа многообразия отношений значений количественных (сырых и латентных) признаков на числовой оси и вычисления функции принадлежности объектов к классам предлагается использовать критерии для поиска оптимальных границ непересекающихся интервалов. Рассмотрим два критерия для разбиения значений признаков на непересекающиеся интервалы по паре классов *Ki*,*Kj*, *i*,*j*∈{1,…,*l*}. Поиску экстремумов этих критериев предшествует упорядочение значений по неубыванию. Для удобства изложения в качестве индексов классов будем использовать *i*=1, *j*=2.

Пусть для значений признака *xc*∊*X*(*n*) в описании объектов *K*1U*K*2 построена упорядоченная по неубыванию последовательность

*r*1,…,*rj*,…,*rh*, *h*=| *K*1U*K*2|. (1)

В качестве границ двух непересекающихся интервалов [π1;π2], (*π*2;*π*3], определяемых по (1), используются π1=*r*1, π2=*rj*, 1<*j*<*h*, π3=*rh*. Интервалы [π1;π2] и (*π*2;*π*3] идентифицируются, соответственно, как первый и второй. Вес признака у объектов классов по (1) вычисляется как максимум произведения внутриклассового сходства и межклассового различия по критерию из [4]

, (2)

где  – количество значений признака *xc* у объектов из класса *Ki* (*K3-i*) в *d*–ом интервале. Множество допустимых значений критерия (2) принадлежит (0;1] и используется для оценки объектов классов на числовой оси. Если в каждом интервале содержатся все значения признака объектов из одного класса, то его вес равен 1.

Граница между классами (порог) для количественного признака *xa* вычисляется как

, (3)

где *b* – ближайшее к π2 значение из интервала (*π*2; *π*3], определяемого по (2). При вычислении порога по (3) не делается никаких предположений о природе среды данных. Значение (2) интерпретируется как мера компактности объектов выборки из двух классов на числовой оси. В данной работе эта мера применяется для оценки наборов исходных признаков по значениям формируемым на них латентных признаков.

Определим ограничения на использование альтернативного (2) критерия, число непересекающихся интервалов для реализации которого изначально неизвестно. Пусть  – число отличных друг от друга значений признака по (1) и *aj*1,…,*ajq* их количество в классе *Kj*, *j*=1,2. В случае равенства

 (4)

не существует интервалов, в границах которых частота встречаемости значений признака у объектов из класса *Kt* будет больше чем частота встречаемости у объектов из класса *K*3-*t*, *t*=1,2.

При отсутствии ограничения (4) для разбиения (1) на множество из *pc* (*pc*≥2) непересекающихся интервалов {[*ru*;*rv*]*i*}, 1≤*u*, *u*≤ *v*≤ *h*, *i*=1,…,*pc* предлагается использовать критерий из [4]. Значения в границах интервала [*ru*;*rv*]*i* при анализе данных могут рассматриваться как градация номинального признака. Считается, что множество чисел, идентифицирующих *pc* градаций номинального признака, всегда можно взаимно–однозначно отобразить в множество {1,...,*pc*}.

Пусть *dtc*(*u*,*v*), *d*3-*t,c*(*u*,*v*) – количество представителей классов *Kt*, *K*3-*t* в интервале [*ru*;*rv*]*i*, *i*∈{1,...,*pc*}. Для рекурсивной процедуры выбора значений *ru*,*rv* используется критерий

 . (5)

Границы первого интервала [*ru*;*rv*]1 на последовательности (1) вычисляются по максимуму критерия (5). Аналогичным образом определяются границы для [*ru*;*rv*]*τ*, *τ*>1 на значениях (1) не вошедших в [*ru*;*rv*]1,…,[*ru*;*rv*]*τ*-1. Критерием останова процедуры служит покрытие всех значений (1) непересекающимися интервалами.

**3. Нелинейные преобразования признаков и вычисление обобщённых оценок объектов**

Суть нелинейных преобразований признаков сводится к замене их исходных значений на значения функции принадлежности объектов к классам. В целях унификации обозначений вместо *dtc*(*u*,*v*), *t*=1,2 для интервала [*ru*;*rv*]μ по *xc*∈*X*(*n*) будем использовать *dtc*(μ). Значение функции принадлежности *fc*(μ) к классу *K*1 по интервалу [*ru*;*rv*]μ (градации μ∈{1,...,*pc*}) вычисляется как

 (6)

Очевидно, что отношение порядка между градациями номинальных признаков не существует. Чтобы определить отношение порядка предлагается использовать замену градаций признаков на значения функции принадлежности объектов к классам. Множеству из *pc*, 2≤*pc*<*h* допустимых значений (градаций) номинального признака *xc*∈*X*(*n*) можно поставить в соответствие числа 1,2,…,*pc*. При вычислении функции принадлежности *fc*(μ) к классу *K*1 по градации μ∈{1,2,…,*pc*} в качестве *d*1*c*(μ) (*d*2*c*(μ)) используется число объектов класса *K*1 (*K*2) со значением μ.

Аналогично (4) определим ограничение для номинальных признаков. Если для *xc*∊*X*(*n*) существует градация μ∈{1,2,…,*pc*}, для которой выполняется

 (7)

то признак с *fc*(μ)=0.5 нельзя использовать для нелинейных преобразований. Обозначим через *Z*, *Z**X*(*n*) – множество признаков, для которых выполнены условия (4) или (7), и через *D*={*i* | *xi*∈*X*(*n*)*\Z*} – множество индексов признаков, которые можно использовать для нелинейных преобразований.

Замена значений признака *xc*∈*X*(*n*)\*Z* у объектов из [*ru*;*rv*]μ или с градацией μ∈{1,2,…,*pc*} на значение (6) в идеале может привести к корректному (без ошибок) разделению объектов *E*0 на классы. Доказательством этому служит следующий пример. Пусть экстремум критерия (2) на (1) меньше 1 и граница по (3) не даёт корректного разделения объектов на классы *K*1 и *K*2. По критерию (5) на (1) получено разбиение на *pc* (*pc*>2) интервалов, в границах которых представлены объекты только одного класса. В этом случае значения функции принадлежности по (6) по каждой градации μ∈{1,2,…,*pc*} признака *xc*∈*X*(*n*)\*Z* для всех объектов из *K*1 будут равны 1, для *K*2 – 0.

Граница между объектами классов по (6) для *xc*∈*X*(*n*)\*Z* определяется как

*Gc*=(*q*1 + *q*2)/2, (8)

где *q*2=max{*fc*(μ)| 0.5 - *fc*(μ)>0, μ=1,...,*pc*}, *q*1=min{*fc*(μ)| 1 - *fc*(μ)<0.5, μ=1,...,*pc*}. Вес признака *xc* для объектов, значения которых представлены через нелинейные преобразования (6), можно вычислить по (2) либо через градации из {1,2} в номинальной шкале. При вычислении значения градации *aic*, *c*∈*D* для объекта *Si*={*xiu*}*u*∈*D* с использованием (8) и учётом шкал измерений рассматривается одно из двух условий: *xic*∈[*ru*;*rv*]μ либо *xic*=μ. Проверка условий необходима для выбора значений функции принадлежности *fc*(μ) для вычисления *aic* как



Обозначим через  – количество значений градации *j*∊{1,2} признака *xc*∈*X*(*n*)\*Z* в описании объектов соответственно класса *K*1 и *K*2. Межклассовое различие по признаку *xc* определяется как величина

 (9)

Степень однородности (мера внутриклассового сходства) β*c* значений градаций признака по классам *K*1, *K*2 вычисляется по формуле:

 (10)

С помощью (9),(10) вес признака *xc*∈ *X(n)\Z* в номинальной шкале аналогично (2) определяется как произведение внутриклассового сходства и межклассового различия

*wс*=β*с*λ*с*. (11)

Множество допустимых значений весов признаков, вычисляемых по (11), принадлежит интервалу (0;1].

Для вычисления обобщённых оценок объектов [3] на *E*0 используются вклады градаций признаков. Вклад градации *j*∈{1,2} признака *xc*∈*X*(*n*)\*Z* определяется как

 (12)

где  – количество значений градации *j* признака *xc* соответственно в классах *K*1 и *K*2, *wс* – вес признака *xc* по (11). Обобщённая оценка объекта *Sr*∈*E*0 по описанию в номинальной шкале измерений *Sr*={*ari*}*i*∈*D* на наборе *X*(*n*)\*Z* и вкладам (12) вычисляется как

 (13)

Базовым понятием для формирования информативных наборов признаков является устойчивость признака. Для вычисления устойчивости используются значения функции принадлежности. Пусть в описании объекта *Sr*∈*K*1U*K2* исходные значения признаков из *X*(*n*)\*Z* заменены на значения функции принадлежности *Sr*={*bri*}*i*∈*D* по (6). Устойчивость признака *xc*∈*X*(*n*)\*Z* вычисляется как

 (14)

Множество допустимых значений (14) принадлежат (0.5;1]. Устойчивость Ψ(*c*)=1, если по границе (8) объекты без ошибок разделяются на классы *K*1 и *K*2.

**4. Группировка признаков по иерархическому агломеративному алгоритму**

Целью использования иерархического агломеративного алгоритма является разбиение множества признаков *X*(*n*)\*Z* на непересекающиеся подмножества по описаниям объектов из *K*1U*K*2 для синтеза латентных признаков. Считается, что для каждого *xi*∈*X*(*n*)\*Z* вычислен вес *wi* по (11) и значения вкладов η*i*(*j*), *j*∈{1,2} по (12). Особенности реализации процесса формирования латентных признаков алгоритмом иерархической агломеративной группировки заключаются в:

– выборе первого признака в группу;

– наборе правил для включения (не включения) признака в группу;

– вычислении значений обобщённых оценок объектов по вкладам (12) признаков из группы.

Обозначим через *P*, *TYPLAM* – множество индексов признаков соответственно из *X*(*n*)\*Z* и формируемых алгоритмом групп. Реализация алгоритма по шагам будет следующей.

Шаг 1. *P*={*i* | *xi*∈*X*(*n*)\*Z* }. *mikdor*=0.

Шаг 2. Вычислить *crit*=10.  *TYPLAM*={*u*}. *mikdor*= *mikdor*+1.

**Цикл** по *t*∊{1,…,*h*} *R*(*St*) *= wu*η*u*(*xtu*). Конец **цикла**; *cr*1=10.

Шаг 3. **Цикл** по *u* ∊*P*. **Цикл** по *t*∊{1,…,*h*} *bt*= *R*(*St*) + *wu*η*u*(*xtu*). Конец **цикла**;

. . *M*1= *M*1 /|*K*1|. *M*2= *M*2 /|*K*2|. θ=0. γ=0.

**Цикл** по *t*∊{1,…,*h*} Если *St* ∊ *K*1, то θ= θ +|*bt* – *M*1|, γ= γ + |*bt* – *M*2|. Иначе θ= θ +|*bt* – *M*2|, γ= γ + |*bt* – *M*1|. Конец **цикла;**

Если θ/γ < *crit*, то *crit*= θ/ γ, *q*=*u*. Конец **цикла;**

Шаг 4. Если *cr*1< *crit*, то *crit*=*cr*1. *P*=*P*/{*q*}. *TYPLAM* = *TYPLAM* U{*q*}. *cr*1=10.

**Цикл** по *t* ∊{1,…,*h*} *R*(*St*)= *R*(*St*) + *wq*η*q*(*xtq*). Конец **цикла;**

Иначе вывод {*R*(*St*)}*t* ∊{1,…,*h*}, *TYPLAM*.

Шаг 5. Если |*P*|≥2, то идти 2; Иначе вывод *mikdor*.

Шаг 6. Конец.

Множество значений {*R*(*St*)}*t*∊{1,…,*h*}, полученное на шаге 4 алгоритма, формируют описания объектов *K*1, *K*2 по набору *Y*(*mikdor*)=(*y*1,…,*ymikdor*) в латентном признаковом пространстве, размерность которого *mikdor*<*n*. Каждому *yi*∊*Y*(*mikdor*) соответствует набор *X*(*ni*)⸦ *X(n)\Z*. Максимальное значение (2) по *yс*∊*Y*(*mikdor*) служит условием выбора набора *X*(*nс*) в качестве информативного на *K*1U*K*2.

**5. Отбор и упорядочивание наборов информативных признаков**

Иерархический агломеративный алгоритм группировки, описанный выше, реализует жадную стратегию выбора латентного признакового пространства. Размерность пространства при таком выборе будет минимальной. Введением и подбором параметра ε (ε>0) в условии *cr*1< *crit* - ε на шаге 4 можно получить размерность выше минимальной.

Как правило, целью использования разных эвристик для отбора информативных признаков является уменьшение комбинаторной сложности вычислений через сокращения числа переборов различных вариантов. Сокращение числа переборов достигается за счёт:

– применения принципа динамического программирования в процессе реализации иерархического агломеративного алгоритма при формировании непересекающихся наборов признаков и синтезе на основе их значений латентных показателей(признаков);

– использования правила отбора информативного набора признаков по максимальному значению меры компактности (2) по латентным показателям.

При отборе единственного набора для всех классов [6] была использована идея ранжирования признаков по значениям их устойчивости (14) по парам (*K*1,*K*2),…,(*K*1,*Kl*). По такому набору признаков вычислялись значения количественной меры здоровья на данных 14 возрастных групп от 20 до 85 лет и выше [7]. Было получено подтверждение того, что с возрастом количество здоровья монотонно уменьшается. Свойство монотонности по единственному набору не распространялось на все пары (*Ki*,*Ki*+1),…,(*Ki*,*Kl*) при *i*∈{2,…,*l* -2}.

Предлагаемая технология отбора информативных признаков по парам классов позволяет ранжировать признаки по каждому набору и исследовать свойства отношений на множестве наборов. Процедура отбора на основе алгоритма из п.4 применима для всех пар классов (*Ki*,*Kj*), *i*<*j*, *i*∈{1,…,*l*-1}.

Пусть как и в п.2 *F*(*i*,*j*) обозначает максимальное значение меры компактности, которое вычисляется по (2) на наборе латентных признаков , где *t* (*t*<*n*) – размерность пространства. Согласно [8] последовательность является одной из 5 стандартных типов закономерностей в интеллектуальном анализе данных. Интерес в данной работе представляет исследование последовательностей значений вида

*F*(*i*,*i+*1),…., *F*(*i*,*l*) (15)

по парам классам (*Ki*,*Ki*+1),…,(*Ki*,*Kl*), *i*=1,…,*l*-2 на наличие монотонности по бинарному отношению *H*, *H*∈{≤,≥}. Последовательность (15) является монотонной если *F*(*i*,*j*) *H F*(*i*,*j+*1)= true Ɐ*j*∈{*i*+1,…,*l* -1}. Если отношение *H* верно для всех последовательностей (15) с *i*=1,…,*l*-2, то оно обладает свойством транзитивности. Свойство монотонности может учитываться при формировании базы знаний предметной области и реализации правил вывода по ней.

Количество признаков, входящих в состав информативного набора, формируемого иерархическим алгоритмом из п.4, может представляться двухзначным числом. В силу особенностей физической памяти человека ему относительно легко провести анализ и принять решение при числе показателей от трех до семи в описании объектов [1]. В форме вычислительного эксперимента доказывается, что для интерпретации результатов отбора и мотивации ЛПР к определённым действиям достаточно использовать первые по порядку сырые признаки, включённые в информативный набор иерархическим алгоритмом.

**6. Вычислительный эксперимент**

Для эксперимента использовались данные о 452 пациентах с аритмией из UCI Machine Learning [9]. Объекты с 2–мя и более пропусками были удалены из выборки. Также из дальнейшего анализа были исключены показатели *рост*, *вес* и признаки, доля ненулевых значений в которых не превышала 2%. По показателю *возраст* выборка была разделена на 5 непересекающихся классов по годам: *K*1(0–35), *K*2(36–45), *K*3(46–55), *K*4(56–65), *K*5(65+). Для разбиения исходного набора признаков *X*(150) на непересекающиеся группы(наборы) использовался алгоритм иерархической агломеративной группировки. По каждой паре классов (*Ki*,*Kj*), *i*<*j*, *i*=1,…,4 определены значения латентных показателей по наборам исходных признаков. Значения меры компактности {*F*(*i*,*j*)}, вычисляемые по (2) на латентных показателях информативных наборов признаков объектов из *Ki*U*Kj*, показаны в табл. 1.

Таблица 1. Значения меры компактности (2) по латентным показателям, сформированных на информативных наборах признаков по парам классам

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Класс | *K*1 | *K*2 | *K*3 | *K*4 | *K*5 |
| *K*1 | 0.0 | 0.9420 | 0.8663 | 0.9561 | 0.9278 |
| *K*2 | 0.0 | 0.0 | 0.8996 | 0.8706 | 0.9237 |
| *K*3 | 0.0 | 0.0 | 0.0 | 0.9532 | 0.9731 |
| *K*4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.9696 |

Значения мер компактности из табл. 1 близкие к 1(разделению классов без ошибок) являются показателями качества отбора информативных признаков. Свойство монотонности по отношению “≤” (см. табл. 1) выполняется по значениям мер компактности (2) на парах (*K*3,*K*4), (*K*3,*K*5).

Для определения значений обобщённых оценок объектов по (13) не существует проблемы проклятия размерности, свойственной алгоритмам, использующим меры расстояния. При вычислении обобщённых оценок (значений латентного признака) по всему набору *X*(150) размерность пространства для описания объектов двух классов снижается до одного. Эффективность такого снижения размерности через значения меры компактности (2) демонстрируются в табл. 2.

Таблица 2. Значения меры компактности (2) по латентному показателю на наборе *X*(150) по парам классов

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Класс | *K*1 | *K*2 | *K*3 | *K*4 | *K*5 |
| *K*1 | 0.0 | 0.6868 | 0.626923 | 0.6854 | 0.7037 |
| *K*2 | 0.0 | 0.0 | 0.5586 | 0.6051 | 0.6574 |
| *K*3 | 0.0 | 0.0 | 0.0 | 0.6225 | 0.8030 |
| *K*4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.8303 |

Из анализа результатов эксперимента по табл.1 и табл. 2 следуют такие выводы:

– существуют подмножества признаков из набора *X*(150), значения меры компактности (2) по обобщённым оценкам на которых выше чем на *X*(150);

– показана эффективность использования иерархического алгоритма группировки для отбора информативных признаков. По всем парам классов (*Ki*,*Kj*), *i*≠*j* значение меры компактности (2) по информативных наборах выше чем на всём наборе.

Эффективность применения информативных наборов для принятия решений зависит от их мощности. Значимость признаков определяется порядком их включения в информативный набор на шаге 3 иерархическим алгоритмом из п.4 по отношению θ/γ (внутриклассовое сходство)/(межклассовое различие). Демонстрация истинности этого утверждения по паре (*K*3, *K*4) показана в табл. 3.

Таблица 3. Процесс формирования информативного набора признаков

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Число признаков в наборе | Отношение  θ/γ | Значение по (2) | Граница между классами (3) | Точность  в % |
| 2 | 0.6490 | 0.3561 | 0.0000 | 68.45 |
| 3 | 0.5524 | 0.4194 | 0.0000 | 76.79 |
| 4 | 0.5250 | 0.4032 | 0.1189 | 75.60 |
| 5 | 0.4710 | 0.5078 | 0.0437 | 82.14 |
| 6 | 0.4111 | 0.5309 | 0.1641 | 83.33 |
| 7 | 0.3723 | 0.5551 | 0.0396 | 84.52 |
| 8 | 0.3589 | 0.5810 | 0.1129 | 85.71 |
| 9 | 0.3440 | 0.6371 | 0.0364 | 88.10 |

Особенностями процесса отбора информативного набора признаков (см. табл. 3) являются:

– уменьшение значения отношения внутриклассового сходства к межклассовому различию при добавлении нового признака в набор;

– граница (3) между классами находится в окрестности нуля и её положение на числовой оси не зависит от изменения абсолютных значений обобщённых оценок объектов;

– при вычислении значений критерия (2) и границы (3) не делается никаких предположений об функции плотности распределения данных.

Как правило, латентный показатель с максимальным значением меры компактности (2) формируется по признакам из первой или второй группы, составы которых получены по иерархическому алгоритму группировки из п.4. С целью сокращения сложности вычислений этим свойством рекомендуется пользоваться при отборе информативных признаков на выборках с большой (20 тыс. и более) размерностью пространства.

**Заключение**

Предложена технология отбора и анализа информативных наборов разнотипных признаков с использованием нелинейных преобразований на основе функций принадлежности и вычислением обобщённых оценок объектов (латентных признаков) по парам непересекающихся классов. Показана эффективность использования алгоритма иерархической группировки для формирования наборов признаков и выбора информативного из них. Показателем эффективности служит мера компактности значений латентного признака в описании объектов двух классов. Целью использования информативных наборов признаков вместо одного является отказ от принципа усреднения, который повышает степень неопределённости при принятии решений.

**Литература**

1. Петровский А.Б., Лобанов В.Н. Многокритериальный выбор в пространстве признаков большой размерности: мультиметодная технология ПАКС-М // Искусственный интеллект и принятие решений 3/2014. С. 92–104.
2. Шумаков В.И., Новосельцев В.Н., Сахаров М.П., Штенголд Е.Ш. Моделирование физиологических систем организма. – М: Медицина. 1971. 352 с.
3. Ignatiev N. A. On Nonlinear Transformations of Features Based on the Functions of Objects Belonging to Classes // Pattern Recognition and Image Analysis. 2021. V. 31. № 2. P. 197–204.
4. Згуральская Е.Н. Устойчивость разбиения данных на интервалы в задачах распознавания и поиск скрытых закономерностей // [Известия Самарского научного центра Российской академии наук](https://elibrary.ru/contents.asp?id=36874471), 2018. Т. 20. [№ 4(3](https://elibrary.ru/contents.asp?id=36874471&selid=36874494)). С. 451–455.
5. Joseph R. (2019, April 23). Ensemble methods: bagging, boosting and stacking. Understanding the key concepts of ensemble learning. <https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205>.
6. Рахимова М. А. О количественной оценке состояния здоровья по возрастным группам // III Международное книжное издание стран Независимых Государств «Лучший молодой ученый-2021». 2021. С.73-77.
7. Данные медицинского обследования <https://www.data.go.kr> (по состоянию на 28 декабря 2020 г.).
8. Дюк В. А. Data Mining – интеллектуальный анализ данных. – <http://www.olap.ru/basic/dm2.asp>.
9. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].

**Игнатьев Николай Александрович.** Национальный Университет Узбекистана им. Мирзо Улугбека, г. Ташкент, Узбекистан. Доктор физико-математических наук, профессор. Количество печатных работ: 123 (в т.ч. 3 монографии). Область научных интересов: разработка алгоритмов интеллектуального анализа данных. e-mail: n\_ignatev@rambler.ru. (Ответственный за переписку).

**Рахимова Мехрбону Акром кизи.** Национальный Университет Узбекистана им. Мирзо Улугбека, г. Ташкент, Узбекистан. Магистр. Количество печатных работ: 20. Область научных интересов: интеллектуальный анализ данных, программирование. e-mail: mehribonu@gmail.com.

**Formation and analysis of sets of informative features of objects by pairs of classes**

N.A. IgnatyevI, M.A. RahimovaI

INational University of Uzbekistan named after Mirzo Ulugbek, Tashkent, Uzbekistan

**Abstract:** The formation of informative feature sets is considered, the selection of which is carried out according to pairs of non-overlapping classes of objects. It is considered that the class indices belong to the set of admissible values of the classification attribute in the ordinal scale. The selection procedure is implemented by the hierarchical agglomerative grouping algorithm by dividing the set of initial features into disjoint groups (feature sets). To reduce the dimension of space in the description of objects, the calculation of latent indicators by groups is used. A measure of the compactness of classes by the latent indicator for assessing a set of features has been determined. The estimation of the set with the maximum value of the compactness measure is a condition for choosing it as informative. The properties of binary relations between classes with respect to the values of the compactness measure are investigated.

**Key words:** nonlinear transformations, membership functions, hierarchical agglomerative grouping, generalized estimates of objects, measure of compactness.

**References**

* + - 1. Petrovskiy A.B., Lobanov V.N. Mnogokriterialniy vibor v prostranstve priznakov bolshoy razmernosti: multimetodnaya texnologiya PAKC-M [Multi-criteria choice in the space of high-dimensional features: multi-method technology PAKS-M] // Iskusstvennyj intellekt i prinyatie reshenij [Artificial Intelligence and Decision Making] 3/2014. 92–104.
      2. Shumakov V.I., Novosel'cev V.N., Saharov M.P., SHtengold E.SH. Modelirovanie fiziologicheskih sistem organizma [Modeling the physiological systems of the body]. – M: Medicina [Medicine]. 1971. 352 p.
      3. Ignatiev N. A. On Nonlinear Transformations of Features Based on the Functions of Objects Belonging to Classes // Pattern Recognition and Image Analysis. 2021. V. 31. № 2. P. 197–204.
      4. Zgural'skaya E.N. Ustojchivost' razbieniya dannyh na intervaly v zadachah raspoznavaniya i poisk skrytyh zakonomernostej [Stability of data partitioning into intervals in recognition problems and search for hidden patterns] // Izvestiya Samarskogo nauchnogo centra Rossijskoj akademii nauk [Bulletin of the Samara Scientific Center of the Russian Academy of Sciences], 2018. T. 20. № 4(3). 451–455.
      5. Joseph R. (2019, April 23). Ensemble methods: bagging, boosting and stacking. Understanding the key concepts of ensemble learning. <https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205>
      6. Raximova M. A. O kolichestvennoy otsenke sostoyaniya zdorovya po vozrastnыm gruppam // III Mejdunarodnoe knijnoe izdanie stran Nezavisimыx Gosudarstv «Luchshiy molodoy uchenny-2021» [III International book edition of the countries of the Commonwealth of Independent States “BEST YOUNG SCIENTIST - 2021 ”]. 2021. 73-77.
      7. Dannye medicinskogo obsledovaniya [Medical examination data]. Available at: https://www.data.go.kr (accessed December 28, 2020).
      8. Дюк В. А. Data Mining – intellektualniy analiz dannix [Data Mining] – <http://www.olap.ru/basic/dm2.asp>.
      9. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml].

**N.A. Ignatyev.** Doctor of Science in physics and mathematics, professor, National University of Uzbekistan, 4 University Str., Tashkent 100174, Uzbekistan; n\_ignatev@rambler.ru (corresponding author).

**M.A. Rakhimova.** National University of Uzbekistan, 4 University Str., Tashkent 100174, Uzbekistan; mehribonu@gmail.com.