Комбинированное использование корреляционных мер в задаче отбора концептов при построении онтологии

А.Ю. Тимофеева[1], Т.В. Авдеенко[1], Е.С. Макарова[1], М.Ш. Муртазина[1]

[1]Новосибирский государственный технический университет, пр. К. Маркса 20, Новосибирск, Россия, 630073

Аннотация. В работе предлагается новый подход к отбору концептов для построения онтологии. Он основан на методе главных компонент, но, в отличие от стандартного подхода, используются не коэффициенты корреляции Пирсона, а другие корреляционные меры. Это связано с тем, что отбор концептов производится по данным о семантической связи между концептами и прецедентами, представленным в виде весовых коэффициентов, принимающих дискретные значения и значительное число нулевых значений. Для таких случаев наиболее подходящим является полихорический коэффициент корреляции, он позволяет выявлять монотонную зависимость по таблицам сопряженности. Однако при определенной структуре таблиц коэффициент ошибочно указывает на тесную связь. Именно эта проблема детально проанализирована, и предложено в проблемных случаях использовать корреляционное отношение. На примере задачи отбора концептов для построения онтологии области ИТ-консультирования показаны преимущества предложенного подхода, состоящие в увеличении процента дисперсии концептов, объясненной выделенными компонентами.

1. Введение

Одно из ключевых направлений развития искусственного интеллекта связывается с переходом от хранения и обработки данных к накоплению и обработке знаний. В этом процессе важная роль отводится онтологии как форме представления знаний. Основными составляющими онтологии выступают понятия (концепты) предметной области. При построении онтологии важно отобрать концепты таким образом, чтобы избежать их избыточности. Эта задача может быть рассмотрена в рамках машинного обучения как выбор подходящего подмножества признаков. Он осуществляется как до начала обучения, так и совместно с решением задачи построения модели.

Существует несколько подходов к решению этой задачи. Подходы на базе фильтров оценивают признаки по индивидуальным характеристикам (прирост информации, статистика хи-квадрат и др.). Примером служит алгоритм отбора Relief [1, 2]. К недостаткам фильтрации можно отнести то, что не учитывается избыточность признаков, не обнаруживается зависимость между признаками. Эти недостатки пытаются устранить путем использования «оберток» – процедур поиска, включающих обучение и оценку модели с помощью потенциального подмножества признаков. При этом используются методы включения и исключения. Однако такие процедуры требуют в идеале перебора всех возможных подмножеств множества признаков, то есть алгоритмы характеризуются экспоненциальной

сложностью по числу признаков. Это, как правило, неприемлемо, и приходится прибегать к «жадным» алгоритмам поиска, которые никогда не пересматривают сделанного ранее выбора.

Другой вариант – это использование регуляризации при построении моделей. Примером может служить регрессия lasso [3], которая стягивает веса одних признаков и обнуляет веса других. Тем самым получается разреженное решение, включающее только существенные признаки. Оценки такой регрессии, однако, не выражаются аналитически, что требует применения численных методов оптимизации. Кроме того решение очень чувствительно к параметру регуляризации, влияющему на степень разреженности решения. Как правило, он выбирается на основе перекрестной проверки, однако, существует пробел в исследованиях, посвященных оптимальному выбору параметра регуляризации.

Наконец, признаковое пространство может быть сокращено путем применения методов классификации и снижения размерности. Самыми популярными подходами здесь остаются метод главных компонент и факторный анализ, использующие разложение ковариационной (корреляционной) матрицы.

Однако при использовании коэффициентов корреляции необходимо учитывать, что данные могут быть разнотипными. Обычно используется коэффициент корреляции Пирсона. Однако он не подходит, если есть бинарные и порядково-категориальные переменные. В работе [4] показано, что когда анализируется валидность конструктов по порядковым данным, измеренным в шкале Ликерта, результаты факторного анализа лучше отражают теоретическую модель, когда факторизация выполняется с использованием полихорических корреляций, а не коэффициентов корреляции Пирсона. Тем не менее полихорический коэффициент корреляции обладает рядом недостатков, в частности, при определенной структуре таблицы сопряженности он ошибочно показывает наличие сильной взаимосвязи. Эта проблема особенно характерна для разреженных таблиц с большим числом нулевых значений. Далее анализируются причины такого поведения коэффициента и предлагаются пути его коррекции за счет комбинированного использования корреляционных мер.

2. Полихорический коэффициент корреляции

Для анализа связей между переменными, которым трудно дать объективную количественную оценку и значения которых представляют собой упорядоченные категории, предназначен полихорический коэффициент корреляции. Его можно использовать и в том случае, если анализируются счетные данные, то есть дискретные, принимающие ограниченное количество числовых значений. Речь может идти и об округленных данных, а также о тех случаях, когда данные определены субъективно и неточно, например, в ходе опросов экспертов.

2.1. Определение

Пусть наблюдаются дискретные признаки x_1 и x_2. Им ставятся в соответствие латентные переменные ξ_1 и ξ_2, они считаются непрерывными. Поскольку совместное распределение латентных переменных ξ_1 и ξ_2 неизвестно, то обычно вводятся априорные предположения. В настоящее время известны работы с использованием скошенных распределений и распределений с тяжелыми «хвостами» [5]. Однако традиционным является предположение о двумерном стандартном нормальном распределении величин ξ_1 и ξ_2 с коэффициентом корреляции ρ, которое было сделано авторами и в рамках данной работы.

Введенные величины x_1 и x_2 определяются на основе группирования латентных переменных ξ_1 и ξ_2, то есть разбиения области их значений на интервалы. Предполагается, что x_1 принимает значения от 1 до n_1, x_2 – от 1 до n_2, где n_1, n_2 – число интервалов разбиения латентных переменных ξ_1, ξ_2, соответственно. Границы этих интервалов α_{i1}, $i = 0,1,\ldots,n_1$, α_{j1}, $j = 0,1,\ldots,n_2$, называются порогами дискретизации. Они неизвестны и $\alpha_{01} = \alpha_{02} = -\infty$, $\alpha_{n_1 1} = \alpha_{n_2 2} = +\infty$. Тогда соотношение между x_k и ξ_k можно записать в виде

$x_k = i$, если $\alpha_{(i-1)k} < \xi_k < \alpha_{ik}$, $i = 1,\ldots,n_k$, $k = 1,2$.

Вероятность реализации значений показателей, измеренных в ходе опроса, определяется как

$$\mathrm{P}(x_k = i) = \mathrm{P}(\alpha_{(i-1)k} < \xi_k < \alpha_{ik}) = \Phi(\alpha_{ik}) - \Phi(\alpha_{(i-1)k}),\ i = 1,\ldots,n_k,\ k = 1,2,$$

где $\Phi(\cdot)$ – функция стандартного нормального распределения.

В выборочном исследовании i-е значение случайной величины x_1 и j-е значение x_2 совместно наблюдаются с некоторой частотой (числом ответов). Совокупность таких частот представляется в виде таблицы сопряженности. После нормирования на объем выборки можно получить значения относительных частот d_{ij}, совокупность которых далее будем называть таблицей частот.

Теоретическая вероятность $p_{ij} = \mathrm{P}(x_1 = i, x_2 = j)$, соответствующая d_{ij}, определяется как

$$\begin{aligned} p_{ij} &= \mathrm{P}(x_1 = i, x_2 = j) = \mathrm{P}(\alpha_{(i-1)1} < \xi_1 < \alpha_{i1}, \alpha_{(j-1)2} < \xi_2 < \alpha_{j2}) = \\ &= \Phi_2(\alpha_{i1}, \alpha_{j2}, \rho) - \Phi_2(\alpha_{(i-1)1}, \alpha_{j2}, \rho) - \Phi_2(\alpha_{i1}, \alpha_{(j-1)2}, \rho) + \Phi_2(\alpha_{(i-1)1}, \alpha_{(j-1)2}, \rho), \end{aligned} \quad (1)$$

где $\Phi_2(z_1, z_2, \rho)$ – функция двумерного стандартного нормального распределения с коэффициентом корреляции ρ между случайными величинами ξ_1 и ξ_2. Коэффициент корреляции ρ в этой модели называется полихорическим коэффициентом корреляции.

2.2. Постановка задачи оценивания

Задача состоит в оценивании параметра ρ на основе значений относительных частот d_{ij}. В настоящем исследовании рассматривается двухшаговый подход [6]. Первым шагом является вычисление границ интервалов α_{ik} как квантилей соответствующих маргинальных распределений. Предположение о двумерном нормальном распределении влечет нормальность маргинальных распределений. Следовательно, оценки порогов дискретизации можно вычислить следующим образом:

$$\hat{\alpha}_{i1} = \Phi^{-1}\left(\sum_{l=1}^{i}\sum_{j=1}^{n_2} d_{lj}\right),\ i = 1,\ldots,n_1 - 1,\ \hat{\alpha}_{j2} = \Phi^{-1}\left(\sum_{l=1}^{j}\sum_{i=1}^{n_1} d_{il}\right),\ j = 1,\ldots,n_2 - 1,$$

где $\Phi^{-1}(z)$ – обратная функция, или квантиль, стандартного нормального распределения. Маргинальные распределения используются также для расчета p_{in_2} и $p_{n_1 j}$ по соотношению (1) в связи с тем, что

$$\Phi_2(\alpha_{i1}, +\infty, \rho) = \Phi(\alpha_{i1}),\ \Phi_2(+\infty, \alpha_{j2}, \rho) = \Phi(\alpha_{j2}).$$

На втором этапе оценки порогов подставляются в (1), и теоретические вероятности рассматриваются как функция от неизвестного параметра ρ.

Классический метод оценивания параметра ρ – метод максимального правдоподобия (ММП). Для совместного распределения дискретных случайных величин x_1 и x_2 в предположении о независимости наблюдений средний логарифм функции правдоподобия [6] вычисляется как

$$\hat{l} = \sum_{i,j \in U} d_{ij} \ln p_{ij}, \quad (2)$$

где конечное множество $U = \{i, j : d_{ij} \neq 0 \ \&\ p_{ij} \neq 0\}$ вводится во избежание бесконечного значения функции $\hat{l}$ при $d_{ij} = p_{ij} = 0$. В (2) каждое значение d_{ij} фиксировано для заданной выборки.

Максимизация соотношения (2) как функции от параметра ρ и дает ММП-оценку полихорического коэффициента корреляции.

2.3. Преимущества и недостатки

Как и обычный коэффициент корреляции Пирсона, полихорический коэффициент принимает значения от –1 до 1. Абсолютные значения коэффициента близкие к единице свидетельствуют о наличии тесной связи между признаками. При этом если коэффициент корреляции Пирсона выявляет только линейную взаимосвязь, то достоинством полихорического коэффициента корреляции является то, что он позволяет обнаружить и любую монотонную зависимость. Кроме того, как и коэффициент корреляции Пирсона, он симметричен, то есть если поменять x_1 и x_2 местами, коэффициент не изменится. Это обеспечивает свойство симметричности матрицы полихорических корреляций.

Вместе с тем достоинство полихорического коэффициента может обернуться существенным недостатком. Так рассмотрим ряд примеров таблиц относительных частот:

$$D_1 = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0 \end{pmatrix},\ D_2 = \begin{pmatrix} 0.74 & 0.01 \\ 0.25 & 0 \end{pmatrix},\ D_3 = \begin{pmatrix} 0.74 & 0.25 \\ 0.01 & 0 \end{pmatrix}.$$

Во всех трех случаях значение полихорического коэффициента равно –1. Тем самым результат никак не будет зависеть от значений ненулевых частот, главное, чтобы $d_{22} = 0$, а остальные частоты оказались ненулевыми. Но если в первом случае еще можно заподозрить наличие некоторой нелинейной зависимости, то в остальных случаях малая относительная частота 0.01 может быть просто следствием наличия аномального наблюдения.

Проблема сохраняется и для таблиц большей размерности, для которых выполняются условия:

$$d_{1i} \neq 0\ \forall i,\ d_{j1} \neq 0\ \forall j,\ d_{kl} = 0,\ \forall k,\ l \neq 1. \quad (3)$$

В этом случае коэффициент будет показывать строгую отрицательную зависимость. Если матрица будет близка к такой структуре, то коэффициент будет близок к –1 и ошибочно укажет на наличие взаимосвязи. Стоит заметить, что аналогичная проблема характерна, например, для коэффициента Юла, выявляющего взаимосвязь между бинарными переменными. Отмечается, что он неустойчив к малым частотам. Однако в научной литературе не предложено подходов к решению этой проблемы, которые можно было напрямую применить в задаче отбора подходящего подмножества признаков.

3. Комбинированное использование корреляционных мер

Очевидно, что если таблица сопряженности имеет структуру, описываемую соотношениями (3), то применение полихорического коэффициента корреляции приводит к некорректным результатам. По этой причине необходимо привлекать другие корреляционные меры, позволяющие выявлять, в том числе, нелинейные взаимосвязи и пригодные для анализа дискретных переменных. При этом они должны быть более чувствительны к ненулевым значением наблюдаемых частот $d_{1i} \neq 0\ \forall i,\ d_{j1} \neq 0\ \forall j$.

Самым простым подходом было бы заменить полихорический коэффициент корреляции на коэффициент корреляции Пирсона в тех случаях, когда первый ложно указывает на сильную взаимосвязь. Такой тривиальный подход также будет анализироваться, но лучше выбрать меру, более подходящую для анализа взаимосвязей по дискретным данным.

Одним из вариантов является полисериальный коэффициент корреляции, предполагающий, что переменная x_1 дискретна, и ей соответствует латентная переменная ξ_1 со стандартным нормальным распределением, а другая переменная x_2 рассматривается как непрерывная и нормально распределенная. Далее считается, что переменная x_2 предварительно стандартизирована.

Согласно [7], логарифмическая функция правдоподобия для совместного распределения случайного вектора (x_1, x_2), построенная по выборке из n пар наблюдений (x_{i1}, x_{i2}), имеет вид

$$\log L = \sum_{i=1}^{n} \log \phi(x_{i2}) + \log \mathrm{P}\left(x_1 = x_{i1} \mid x_2 = x_{i2}\right), \quad (4)$$

где $\phi(\cdot)$ – функция плотности стандартного нормального распределения.

Условное распределение ξ_1 при заданном $x_2 = x_{i2}$ является нормальным со средним ρx_{i2} и дисперсией $\left(1-\rho^2\right)$. Тогда если $x_{i1} = j$ с категориями $j = 1,\ldots,n_1$, то условная вероятность вычисляется как

$$\mathrm{P}\left(x_1 = j \mid x_2 = x_{i2}\right) = \Phi\left(\frac{\alpha_{j1} - \rho x_{i2}}{\sqrt{1-\rho^2}}\right) - \Phi\left(\frac{\alpha_{(j-1)1} - \rho x_{i2}}{\sqrt{1-\rho^2}}\right).$$

При применении двухшагового подхода оценки порогов дискретизации находятся по формуле:

$$\hat{\alpha}_{j1} = \Phi^{-1}\left(\sum_{l=1}^{j} d_l\right),\ j = 1,\ldots,n_1 - 1,$$

где d_l – это относительная частота, или доля случаев, в которых наблюдается категория l.

В результате максимизации логарифмической функции правдоподобия (4) по аргументу ρ получается оценка полисериальной корреляции. Ясно, что полисериальный коэффициент не симметричен, поэтому здесь для оценки корреляции важно, какая из переменных предполагается непрерывной, а какая – дискретной.

Предположительно, полисериальному коэффициенту корреляции могут быть свойственны те же недостатки, что и полихорическому. Поэтому дополнительно рассмотрим корреляционное отношение случайной величины Y по случайной величине X, определяемое как

$$\eta^2_{Y|X} = 1 - \frac{\overline{D}_{Y|X}}{D_Y}, \tag{5}$$

где $\overline{D}_{Y|X}$ – среднее значение условной дисперсии случайной величины Y при условии X, D_Y – безусловная дисперсия случайной величины Y. Из соотношения (5) очевидно, что корреляционное отношение всегда неотрицательно. Нулевое значение свидетельствует об отсутствии связи. Для сопоставления с коэффициентами корреляции лучше рассматривать величину $\eta_{Y|X}$ или $\eta_{X|Y}$. Как и полисериальный коэффициент, корреляционное отношение несимметрично.

Для анализа возможностей комбинированного использования корреляционных мер авторами реализован расчет полихорического, полисериального коэффициента корреляции и корреляционного отношения в среде R с использованием стандартного набора модулей. Для этого написано несколько пользовательских функций.

4. Практическое применение в задаче отбора концептов онтологии

При построении онтологии важно сформулировать ее концепты так, чтобы избежать избыточности. Тем самым возникает задача отбора подходящих концептов на основе их семантических связей с прецедентами.

Теснота семантической связи определяется некоторыми весами, принимающими значения от 0 до 1. Как правило, веса назначаются экспертным путем, поэтому принимают дискретные значения (например, округлены до десятых).

В эмпирическом исследовании использовались данные о семантической связи прецедентов и концептов онтологии предметной области ИТ-консультирования [8]. Данные содержат 120 прецедентов и 20 концептов. Сначала построена матрица полихорических корреляций между всеми концептами. Всего матрица (нижний треугольник) содержит 190 парных коэффициентов корреляции. В результате обнаружено, что 99 коэффициентов (около половины) близки к –1. Следует отметить, что в проблемных случаях при численной оптимизации функции (2) не всегда значение оценки в точности равно –1, поскольку близкие к –1 дают приблизительно равное значение целевой функции.

Для сопоставления рассчитаны парные коэффициенты корреляции Пирсона r_{xy}. Результаты представлены в виде корреляционного поля на рисунке 1. Здесь и далее (рисунки 2–4) линия соответствует ситуации равенства коэффициентов корреляции, то есть для рисунка 1 – это график уравнения $r_{xy} = \rho$.

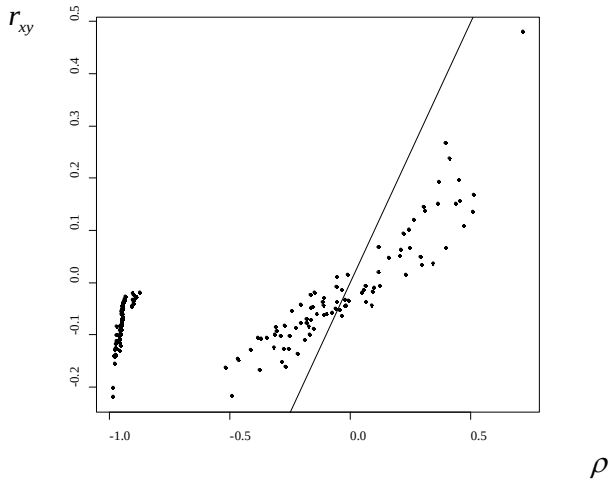

Рисунок 1. Взаимосвязь коэффициентов корреляции полихорического и Пирсона.

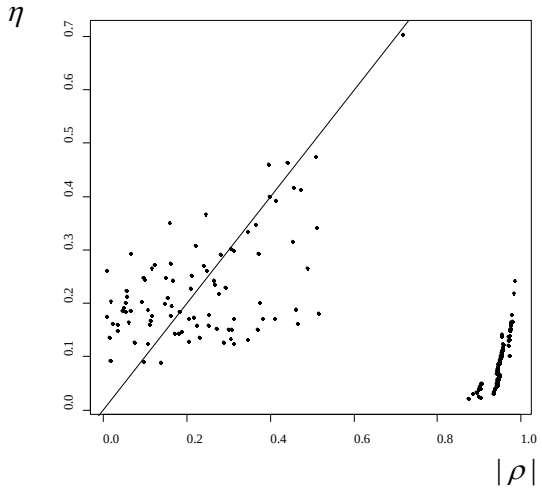

Рисунок 2. Взаимосвязь полихорического коэффициента корреляции и корреляционного отношения.

Из рисунка 1 хорошо видно, что в большинстве случаев полихорический коэффициент корреляции указывает на более тесную связь между концептами, чем коэффициент корреляции Пирсона. Однако хорошо видны и проблемные моменты – близкие к –1 значения полихорического коэффициента. В этих случаях значения коэффициента корреляции Пирсона варьируются от 0 до –0.2, что свидетельствует о довольно слабой связи. Однако коэффициент корреляции Пирсона не выявляет нелинейной зависимости, следовательно, может занизить степень взаимосвязи.

Поскольку как корреляционное отношение, так и полисериальный коэффициент несимметричен, то далее рассчитывалось η как среднее значение между $\eta_{Y|X}$ и $\eta_{X|Y}$, а также ρ_s как среднее между полисериальными коэффициентами, построенными в предположении, что первая из переменных непрерывна, а вторая дискретна, и наоборот.

Как было отмечено выше, корреляционное отношение не указывает направление взаимосвязи, поскольку принимает только неотрицательные значения. По этой причине его корректнее сравнивать с коэффициентами корреляции, взятыми по модулю. Так на рисунке 2 его значения сопоставляется с абсолютными значениями полихорического коэффициента корреляции. Видно, что в ряде случаев более тесную связь показывает корреляционное отношение, а в других – полихорический коэффициент. Видны и проблемные ситуации, в этих случаях корреляционное отношение принимает значения близкие к абсолютным значениям коэффициента корреляции Пирсона, и указывает на слабую связь. Но в целом значения корреляционного отношения η, согласно его свойствам, всегда больше или равны $|r_{xy}|$, поэтому лучше использовать корреляционное отношение. При этом для того чтобы учесть направление взаимосвязи, нужно взять знак полихорического коэффициента корреляции.

Если сравнивать полихорические и полисериальные коэффициенты корреляции (рисунок 3), то в большинстве случаев (77 коэффициентов из 91, не относящихся к проблемным) полихорические коэффициенты указывают на более тесную связь, чем полисериальные. Тем самым полисериальные коэффициенты систематически занижают тесноту связи. Чего не

скажешь о корреляционном отношении: из 91 непроблемных коэффициентов только 44 полихорических коэффициента корреляции по модулю больше корреляционного отношения.

При этом в проблемных случаях полисериальный коэффициент показывает более тесную связь между концептами, поскольку принимает значения от –0.6 до –0.2. Однако это свидетельствует, скорее о том, что этот коэффициент также негативно реагирует на определенную структуру таблиц сопряженности. Это хорошо видно и из рисунка 4, где сопоставляются абсолютные значения полисериального коэффициента и корреляционного отношения. Серым цветом выделены ситуации, в которых значения полихорического коэффициента близко к –1. Они, очевидно, выделяются на фоне остальных точек на графике.

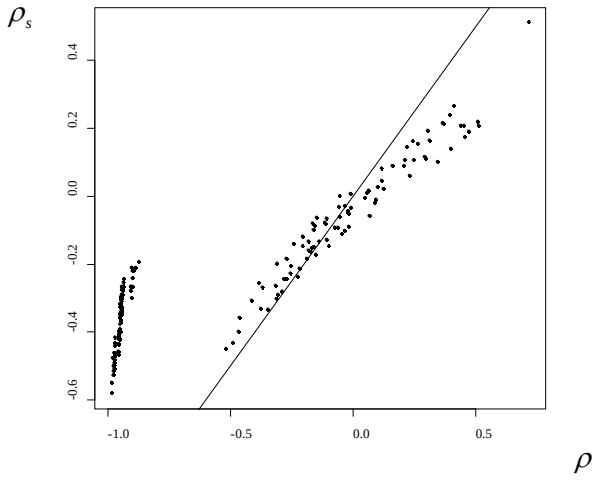

Рисунок 3. Взаимосвязь полихорического и полисериального коэффициентов корреляции.

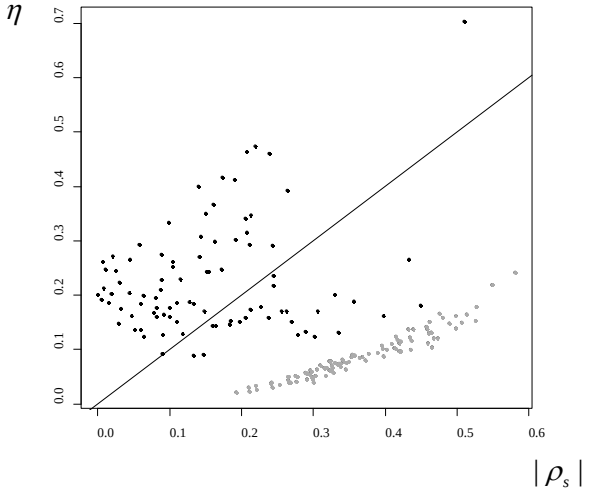

Рисунок 4. Взаимосвязь полисериального коэффициента корреляции и корреляционного отношения.

Таким образом, предлагается подход к отбору подходящих концептов для онтологии, состоящий из следующих этапов.

Этап 1. Вычисление полихорических коэффициентов корреляции ρ.

Этап 2. Идентификация проблемных ситуаций по таблицам частот, для которых выполняются условия (3), а также по значениям полихорических коэффициентов корреляции, близких к –1.

Этап 3. Замена полихорических коэффициентов корреляции в проблемных ситуациях, выявленных на этапе 2, на значения корреляционных отношений η, вычисленных как среднее между $\eta_{Y|X}$ и $\eta_{X|Y}$, взятых со знаком $\text{sign}(\rho)$.

Этап 4. На основе итоговой корреляционной матрицы, состоящей из полихорических коэффициентов и корреляционных отношений, реализация метода главных компонент, вычисление нагрузок на основные компоненты и выделение блоков взаимосвязанных концептов.

Преимущества использования предложенного подхода по сравнению со стандартным (вычисление корреляции Пирсона) должно состоять в увеличении процента дисперсии концептов, объясненной выделенными компонентами. В конечном итоге это позволяет разбить концепты на меньшее число групп, взаимосвязи внутри которых более тесные.

Проверим, что в действительности достигается такой эффект. На рассмотренном примере анализа концептов онтологии выделим 5 компонент с помощью стандартного и предложенного подхода. Результаты представлены в таблице 1. Абсолютные значения нагрузок отражают тесноту связи между концептами и главными компонентами. Для удобства восприятия значения, не превышающие по модулю 0.3, в таблице 1 не приведены. Слабая связь концептов (например, «отпуск» при стандартном подходе) со всеми пятью компонентами указывает на то, что такие концепты не удалось включить в выделенные группы.

Таблица 1. Нагрузки на главные компоненты и накопленный процент объясненной дисперсии.

Концепт	Стандартный подход					Предложенный подход				
	1	2	3	4	5	1	2	3	4	5
Прием		0.343					0.423			
Увольнение		0.316			0.337		0.365			
Отпуск						-0.367				
Больничный		-0.362				-0.300				
Табель	0.370	-0.340				-0.321				
Отчеты		0.520					0.514			
Аванс				-0.392						0.438
Начисления основные					-0.358				0.563	
Средний заработок					0.356				-0.408	
Удержание			0.355	-0.362						0.364
Выплата				-0.488						0.440
Перерасчет			0.514					-0.414		
2.НДФЛ	-0.535					0.489				
6.НДФЛ	-0.490		-0.366			0.424				
Страховые взносы			0.338	0.347	-0.302			-0.461		
Прочие налоги								-0.335		
Проводки								-0.314	0.312	
Накопленная объясненная дисперсия, %	9.7	18.1	25.9	32.7	**38.9**	14.7	27.0	37.7	46.6	**55.1**

Как видно, из результатов таблицы 1, благодаря предложенному подходу существенно увеличивается процент объясненной дисперсии. Так при стандартном подходе пять выделенных компонент суммарно объясняют только 38.9% исходной вариации концептов, тогда как предложенный подход позволяет объяснить 55.1% дисперсии. Кроме того в пять выделенных групп удалось включить большее число концептов, дополнительно включены «отпуск», «прочие налоги» и «проводки». Тем самым желаемый эффект достигается.

5. Выводы

Таким образом, в ходе отбора концептов для построения онтологии на основе их семантических связей с прецедентами применение метода главных компонент требует выбора подходящих корреляционных мер. В силу особенностей данных, представляющих весовые коэффициенты и принимающих дискретные значения, предложено использовать полихорический коэффициент корреляции. Однако, как оказалось, он дает некорректные результаты при определенной структуре таблиц сопряженности. При этом в проведенном эмпирическом исследовании онтологии предметной области ИТ-консультирования такая структура встречается достаточно часто (в половине случаев). Поэтому предлагается в проблемных ситуациях заменять полихорический коэффициент корреляции на корреляционное отношение, выявляющее нелинейные связи и применимое для дискретных данных. Такое комбинированное использование корреляционных мер в конечном итоге позволяет увеличить процент объясненной дисперсии при применении метода главных компонент.

6. Благодарности

Работа поддержана грантом Министерства образования и науки РФ в рамках проектной части государственного задания, проект № 2.2327.2017/4.6 «Интеграция моделей представления знаний на основе интеллектуального анализа больших данных для поддержки принятия решений в области программной инженерии».

7. Литература

[1] Kira, K. A practical approach to feature selection / K. Kira, L.A. Rendell // Proceedings of the ninth international workshop on Machine learning, 1992. – P. 249-256.
[2] Robnik-Sikonja, M. Theoretical and empirical analysis of ReliefF and RReliefF / M. Robnik-Sikonja, I. Kononenko // Machine learning. – 2003. – Vol. 53(1-2). – P. 23-69.
[3] Tibshirani, R. Regression shrinkage and selection via the lasso / R. Tibshirani // Journal of the Royal Statistical Society. Series B (Methodological). – 1996. – Vol. 1. – P. 267-288.
[4] Holgado-Tello, F.P. Polychoric versus Pearson correlations in exploratory and confirmatory factor analysis of ordinal variables / F.P. Holgado-Tello // Quality & Quantity. – 2010. – Vol. 44(1). – P. 153-166.
[5] Uebersax, J.S. A Latent Trait Finite Mixture Model for the Analysis of Rating Agreement / J.S. Uebersax, W.M. Grove // Biometrics. – 1993. –Vol. 49. – P. 823-835.
[6] Olsson, U. Maximum Likelihood Estimation of the Polychoric Correlation Coefficient / U. Olsson // Psychometrica. – 1979. – Vol. 44. – P. 443-460.
[7] Drasgow, F. Encyclopedia of statistical sciences / F. Drasgow // Polychoric and polyserial correlations. – John Wiley & Sons, 1988.
[8] Авдеенко, Т.В. Система поддержки принятия решений в IT-подразделениях на основе интеграции прецедентного подхода и онтологии / Т.В. Авдеенко, Е.С. Макарова // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2017. – Т. 3. – С. 85-99.

Combined use of correlation measures in the task of selecting concepts in the construction of ontology

A.Yu. Timofeeva[1], T.V. Avdeenko[1], E.S. Makarova[1], M.Sh. Murtazina[1]

[1]Novosibirsk State Technical University, K. Marx Ave. 20, Novosibirsk, Russia, 630073

Abstract. The paper suggests a new approach to the selection of concepts for the construction of ontology. It is based on the principal component analysis, but, unlike the standard approach, not Pearson correlation coefficients, but other correlation measures are used. This is due to the fact that the selection of concepts is based on data on the semantic connection between concepts and cases, which are represented in the form of weight coefficients that take discrete values and a significant number of zero values. For such cases, the most suitable is the polychoric correlation coefficient. It allows one to detect a monotonous dependence on the contingency table. However, for a certain table structure, the coefficient erroneously indicates a close relationship. It is this problem that has been analysed in detail, and it has been suggested to use the correlation ratio in problem cases. Using the example of the problem of selecting concepts for constructing the ontology of the IT consulting area, the advantages of the proposed approach are shown, consisting in increasing the percentage of variance of concepts explained by the principal components.

Keywords: correlation, principal component analysis, ontology, concept, polychoric correlation coefficient, correlation ratio.