Extracting information from rare events in regression analysis

Dushyn, Oleksiy and Dushyn, Borys

Maastricht University

3 February 2024

Extracting information from rare events in regression analysis

Dushyn, Oleksiy and Dushyn, Borys

Inspired at Maastricht University

24 February 2024

Extracting information from rare events in regression analysis

Abstract

This paper investigated an important practical problem of extracting information from rare events in sparse and high-dimensional data while building a linear regression model. It analyzes the advantages and the limitations of the different linear regression method used for high-dimensional problems. Main known methods were selected and tested on the real Tripadvisor.com dataset. The results of this research show the importance of the data aggregation based on hierarchical clustering. It allows extracting information from rare features by aggregating them according the clustering tree. Comparative analyses of main different linear regression methods that use clustering aggregation were done.

Keywords: rare events, regression analysis, sparse data, high-dimensional data, Lasso, Ridge, ElasticNet, rare methods, text mining, semantic aggregation, hierarchical clustering, vector word representation.

Jel classification: C51, C63, C87.

Извлечение информации из редких событий в регрессионном анализе

Рассматривается проблема оценки влияния на отклик редких событий в регрессионном анализе на разреженных данных большой размерности. Обычно такие события считаются шумовыми и игнорируются. В то же время такие события могут иметь сильное влияние на отклик. Агрегирование предикторов на основе дерева их кластеризации позволяет извлекать информацию и из редких событий путем включения редких предикторов в регрессионную модель. Проведен сравнительный анализ различных методов подбора линейной регрессии, использующих агрегирование предикторов.

Ключевые слова: редкие события, регрессионный анализ, разреженные данные, анализ многомерных данных, методы *LASSO, гребневый (ridge), ElasticNet, TASSO*, *rare*, *text mining*, агрегированные (композиционные) данные, семантическое агрегирование, адаптивное агрегирование, иерархическая кластеризация, векторное представление слов.

Jel classification: *C51, C63, C87.*

Введение

Современные массивы данных обычно имеют большую размерность с большим количеством различных характеристик. Нередко количество характеристик значительно превышает число наблюдений. Например, проект Атласа раковых генов (TCGA) уже содержит более 3 млн. мутаций и более 20 тыс. генов (Weinstein et al., 2013). Однако в отличие от большого количества возможных генных мутаций, число раковых пациентов с результатами генных исследований сравнительно невелико. Аналогичная ситуация наблюдается при микробиологических исследованиях (Wang, Zhao, 2017) и при разработке и опробовании новых лекарств и вакцин (Alan Talevi et al., 2020). Такая особенность делает невозможным использования классических подходов статистического анализа данных с помощью линейной регрессии, так как существует бесконечное множество решений, для которых значение целевой функции равно нулю.

Кроме того, высокая разреженность наблюдений в самой выборке еще более затрудняет анализ. Например, если исследуемая характеристика представляет количество определённых событий, то часто возникает ситуация, в которой только незначительная часть событий представлена во всех наблюдениях, а основная их масса присутствует только в одном или в очень малом числе наблюдений. Такие события естественно называют редкими и в регрессионных моделях им соответствуют сильно разреженные предикторы, которые также будем называть редкими. Например, в задачах машинного анализа текстов документов или проверке патентов на изобретение заключение о том или ином документе базируется на используемых словах. Документ представляется в виде вектора с размерностью равной числу слов в лексиконе языка, где каждая компонента вектора является количеством раз использования соответствующего слова в документе (Package ‘tm’, 2023). Часто, многие слова используются

только один раз во всём документе или вообще не используются. Еще пример — прогнозирование поведения пользователя интернета на основании посещённых ранее сайтов. Только малая часть сайтов посещается большим числом пользователей и подавляющее число сайтов посещается не более одного раза. Аналогично — анализ микробиологических сред организмов, которые могут содержать уникальные биологические единицы, существенные для всего организма.

Традиционный регрессионный анализ на основании редких предикторов невозможен так как, частота их проявления пренебрежительно мала в сравнении с размером выборки. Такие предикторы обычно рассматривают как неинформативные (шумовые) и просто игнорируют. Нередко это приводит к игнорированию до 98% исходных данных (Yang, Pederson, 1997) и, как следствие, к высокой ошибке прогнозирования. В то же время такие предикторы могут иметь сильное влияние на отклик. Одним из примеров подобной ситуации является задача определения рейтинга отеля на основании текстовых отзывов постояльцев. Авторы отзывов используют большое количество различных слов, каждое из которых может находиться не более, чем в одном из отзывов. Между тем, каждое слово имеет определённый смысл и может содержать важную информацию о том, почему постоялец дал отелю ту или иную оценку, что ему понравилось, и что нет. Из-за разреженности слов в отзывах традиционные методы регрессионного анализа не в состоянии определить силу влияния редких слов на конечную оценку рейтинга отеля. Однако, естественно предположить, что правильный учёт редких слов в регрессионной модели может повысить точность модели и качество прогнозов.

Целью данной работы является сравнительный анализ основных современных методов обоснованного включения редких предикторов в регрессионную модель. Основной исследовательский вопрос заключается в выборе наиболее эффективного метода позволяющего извлекать информацию из редких событий. Каждый метод оценивается по следующим двум основным параметрам: по среднеквадратичной ошибке прогноза и по количеству редких предикторов, включенных в модель. В первой части работы представляется обзор состояния проблемы на текущий момент. Во второй описываются математические модели и особенности применения основных методов регрессионного анализа на разреженных данных большой размерности. В третьей рассматриваются основные подходы агрегирования предикторов и анализируются методы регрессионного анализа с предварительным и с адаптивным агрегированием предикторов на основе их иерархической кластеризации. В четвёртой части представлены результаты расчёта на реальных данных. В выводах подводятся общие итоги работы и даются рекомендации по дальнейшим исследованиям.

1. Обзор состояния проблемы

Редкость предикторов, разреженность данных, превышение количества характеристик числа наблюдений обычно взаимосвязаны. Задачи, в которых количество характеристик превышает число наблюдений, относятся к классу некорректных задач. На данный момент для некорректных задач существуют три основные метода регрессионного анализа на разреженных данных большой размерности, а именно, гребневый метод, именуемый в иностранных источниках как *ridge* (Тихонов, 1963, Hoerl, Kennard, 1970 и др.), метод *LASSO* (Least Absolute Shrinkage and Selection Operator — оператор наименьшей абсолютной величины и отбора, Tibshirani, 1996, Hastie et al., 2015 и др.), и метод *ElasticNet*

(линейная комбинация *гребневого* метода и метода *LASSO,* Hui Zou, Trevor Hastie, 2005 и др.). Однако при прямом применении все эти методы не в состоянии различать между собой редкие значащие и шумовые предикторы. В методе *LASSO* это приводит к обнулению коэффициентов регрессионной модели для редких значащих предикторов, которые могут иметь сильное влияние на отклик. А в гребневом методе это приводит к увеличению ошибки прогнозирования при включении в модель почти всех предикторов. Метод *ElasticNet* занимает промежуточное положение между первыми двумя методами, обладая как их преимуществами, так и недостатками.

В общем случае для преодоления трудности извлечения информации из редких событий делается преобразование матрицы значений предикторов с целью уменьшения ее разреженности. Такое преобразование может быть предварительным (до начала построения регрессионной модели — uncontrolled aggregation) и адаптивным (в процессе построение регрессионной модели — controlled aggregation). Наиболее распространенны следующие два метода предварительного преобразования.

Первый метод состоит в пороговом отбрасывании предикторов: предикторы, на которых значение функции плотности предиктора меньше определенного порога, отбрасываются. Например, в широко используемом R-пакете программ анализа текстов *tm* (Ingo Feinerer et al., 2023) для этого используется функция *removeSparseTerms.* Такое отбрасывание уменьшает размерность задачи, облегчает расчеты, но не позволяет учитывать редкие предикторы. Второй метод предварительного преобразования состоит в агрегировании предикторов на основе априорного представления о сходстве их влияния на отклик. При этом вначале определяются дополнительные характеристики предикторов, которые могут отражать их влияние на отклик. Затем делается агрегирование предикторов на основе этих характеристик и для агрегированных предикторов подбирается регрессионная модель с использование известных методов регрессионного анализа. При агрегировании важный редкий предиктор попадает в регрессионную модель за счет суммирования значений всех предикторов, которые входят в один агрегированный предиктор (например, Wang, Zhao, 2017). Такой агрегированный предиктор может уже не воспринимается как редкий и может получить ненулевое значение коэффициента в регрессионной модели. Модель на агрегированных предикторах пересчитывают в модель на исходных предикторах, в которой коэффициент при каждом исходном предикторе принимается равным коэффициенту при агрегированном предикторе, содержащем этот исходный предиктор.

Важно, чтобы предикторы, входящие в один агрегированный предиктор, были достаточно «близкими» друг к другу по влиянию на отклик. Такую «близость» для редких предикторов невозможно оценить по обучающим данным. Поэтому для целей агрегирования используется дополнительная информация о предикторах, которая должна позволять оценивать априори сходство их влияния на отклик.

В работе (Yan, Bien, 2018) предложено агрегировать предикторы на основе их иерархической кластеризации. Их расчеты на реальных данных большой размерности и разреженности с использованием метода *LASSO* показали, что включение редких предикторов в регрессионную модель с помощью предварительного агрегирования уменьшает среднеквадратичную ошибку прогноза (*mspe*): 1.36% при 1%-м размере обучающего множества (табл.1). Эффект агрегирования уменьшается с ростом размера обучающего множества.

Для повышения эффективности агрегирования в работе (Yan, Bien, 2018) использовано адаптивное агрегирование, при котором объединение предикторов делается параллельно процессу подбора регрессионной модели. Разработанный *LASSO*-подобный метод *rare* (редкий) показал лучшие результаты, чем метод *LASSO* с предварительным агрегированием (снижение *mspe* в сравнении с отсутствием агрегирования 2.68% при 1%-м размере обучающего множества (табл.1)). Однако, авторы не показали явным образом, что повышение эффективности было достигнуто за счет иного включения в регрессионную модель редких предикторов, а не за счет других факторов, например, иной реализации метода LASSO.

Для подтверждения более высокой эффективности адаптивного агрегирования в сравнении с предварительным необходимы дополнительные расчеты методом *rare* и основными современными методами регрессионного анализа: МНК (метод наименьших квадратов), LASSO, *гребневый* и *ElasticNet.* Для обеспечения условия сопоставимости с методом *rare* испольльзуются кластеризационные деревья с одинаковой топологией. Основными показателями сравнения выбраны среднеквадратичная ошибка прогноза и количество редких предикторов, включенных в модель. В качестве дополнительных показатели — время вычислений и объем занимаемой памяти.

2. Основные современные методы регрессионного анализа

В настоящее время существуют четыре главных метода линейной регрессии. Однако особенности методов влияют на конечный результат построения регрессионной модели, включающей редкие предикторы. С этой точки зрения кратко рассмотрим современные методы регрессионного анализа.

Классический метод наименьших квадратов (МНК). В этом методе коэффициенты $\beta_0, \beta_1, ..., \beta_p$ линейной регрессии $y = \beta_0 - \sum_{j=1}^{p} \beta_j x_j$ определяются из условия минимума среднеквадратичной ошибки

$$\min_{\beta_0, \beta_1, ..., \beta_p} \left(\frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right),$$

где n — размер выборки данных,

p — количество предикторов,

y_i — значение отклика i-ом испытании,

x_{ij} — значение предиктора j в i-ом испытании;

или в матричном виде после нормализации исходных данных

$$\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \| \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} \|_2^2 \right), \quad (1)$$

где $\boldsymbol{X} = \left(x_{ij} \right)_{i=1,...,n;\, j=1,...,p}$; $\boldsymbol{\beta} = \left(\beta_1, ..., \beta_p \right)^T$

Задача (1) при условии невырожденности матрицы $\boldsymbol{X} = \left(x_{ij} \right)_{i=1,...,n;\, j=1,...,p}$ имеет единственное решение. Однако, существует ряд причин, по которым МНК имеет ограниченное применение на разреженных данных большой размерности.

Во-первых, это *точность (аккуратность) модели МНК* (оценивается среднеквадратичной ошибкой прогноза). Эта точность существенно снижается, если тренировочные или тестовые данные содержат выбросы. Выбросы необходимо выявлять и удалять до начала регрессионного анализа. Но при этом все редкие значения считаются выбросами и важные из них по влиянию на отклик тоже удаляются.

Кроме того, *интерпретируемость МНК модели* может быть неудовлетворительной. Часто необходимо выделить небольшое подмножество предикторов, которые проявляют наибольшее влияние на отклик. Но МНК использует все значения предикторов из обучающего множества и метод работает некорректно, если есть корреляция между предикторами. В МНК отсутствует какая-либо селекция предикторов и выбор из них наиболее важных.

И наконец, МНК не способен идентифицировать *приемлемое решение при* $p > n$ (некорректная постановка задачи (1)), когда существует бесконечное множество решений с минимальным (нулевым) значением целевой функции.

Гребневый метод. Еще в прошлом веке Тихонов А. Н. (Тихонов, 1963) предложил для оптимизационных задач большой размерности метод регуляризации и приближения решения. В применении к задаче линейной регрессии этот метод дает функционал решения в виде суммы среднеквадратичной ошибки и регулирующей добавки, равной взвешенной сумме квадратов коэффициентов $\beta_1,...,\beta_p$:

$$\hat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{arg}\left(\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right) \right) \tag{2}$$

где $\|.\|_2$ — l_2 норма вектора;

$\lambda > 0$ — регулирующий параметр, оптимальное значение $\hat{\lambda}$ которого может быть определено по критерию среднеквадратичной ошибки:

$$\hat{\lambda} = arg\left(\min_{\lambda} \left(\frac{1}{2n} \left\| \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda) \right\|_2^2 \right) \right).$$

Однако в этом случае при $p \geq n$ возникает переподгонка модели. Вследствие чего ухудшаются прогнозные способности модели. Для предотвращения этой ситуации применяют кросс-валидацию (перекрестную проверку), которая является одним из методов повторной выборки, направленным на повышение прогнозной способности модели работать на новых данных. Наиболее распространенным методом кросс-валидации является K -блочная кросс-валидация (K -fold cross-validation, например, Berrar D., 2018, Rukshan Manorathna, 2021). В этом методе для λ задается интервал изменения и количество равномерно распределенных в нем значений. Обучающее множество $\{\boldsymbol{X}, \boldsymbol{Y}\}$ делится на K равных (почти равных) по количеству данных непересекающихся блоков $\{\boldsymbol{X}, \boldsymbol{Y}\} = \bigcup_{k=1}^{K} \{\boldsymbol{X}_k, \boldsymbol{Y}_k\}$. После этого для каждого λ делается K расчетов $\hat{\boldsymbol{\beta}}(k, \lambda)$, $k = 1,...,K$:

$$\hat{\boldsymbol{\beta}}(k, \lambda) = arg\left(\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \|\boldsymbol{Y}_{k,train} - \boldsymbol{X}_{k,train}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right) \right),$$

где $\{\boldsymbol{X}_{k,train}, \boldsymbol{Y}_{k,train}\} = \{\boldsymbol{X} \setminus \boldsymbol{X}_k, \boldsymbol{Y} \setminus \boldsymbol{Y}_k\}$ — k -е тренировочное множество, $k = 1,...,K$.

Оптимальное значение $\hat{\lambda}_{CV}$, определенное с помощью кросс-валидации, получают по усредненной среднеквадратичной ошибке моделей с коэффициентами $\hat{\boldsymbol{\beta}}(k,\lambda)$ на соответствующих тестовых множествах:

$$\hat{\lambda}_{CV} = arg(\min_{\lambda}\left(\frac{1}{K}\sum_{k=1}^{K}\left(\frac{1}{2n}\left\|\boldsymbol{Y}_{k,test} - \boldsymbol{X}_{k,test}\hat{\boldsymbol{\beta}}(k,\lambda)\right\|_2^2\right)\right),$$

где $\{\boldsymbol{Y}_{k,test}, \boldsymbol{X}_{k,test}\} = \{\boldsymbol{X}_k, \boldsymbol{Y}_k\}$ — k-е тестовое множество, $k = 1,...,K$.

Оптимальной считается модель с коэффициентами

$$\hat{\boldsymbol{\beta}}(\hat{\lambda}_{CV}) = arg\left(\min_{\boldsymbol{\beta}}\left(\frac{1}{2n}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \hat{\lambda}_{CV}\|\boldsymbol{\beta}\|_2^2\right)\right).$$

Во многих программных пакетах регрессионного анализа $\hat{\boldsymbol{\beta}}(\lambda)$ определяется с использованием кросс-валидации (например, функция cv.glmnet() R-пакета glmnet).

В настоящее время метод Тихонова А. Н. в линейном регрессионном анализе известен под названием *метода ридж, (ridge)* (Hastie et al., 2015).

Гребневый метод стремится приблизить коэффициенты регрессии к нулю, но не присваивает им нулевые значения. Он отбирает в регрессионную модель почти все предикторы, включая и редкие, может давать более точные прогнозы, но не делат селекцию предикторов так как значения коэффициентов получаются приблизительно равными между собой. Рис.1 показывает пример вычислительных путей коэффициентов и среднеквадратичной ошибки регрессии (MSE) на тестовых множествах в процессе подбора значения регулирующего параметра λ_{CV}.

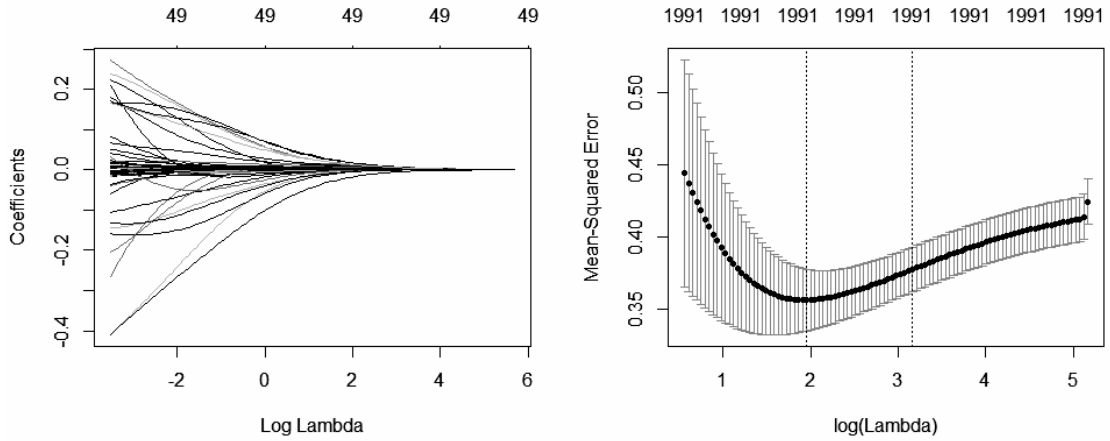

Рис. 1. Изменения величины коэффициентов регрессии и средней квадратичной ошибки (MSE) для гребневого метода в зависимости от значения регулирующего параметра λ. На правом рисунке точки показывают значение средней квадратичной ошибки для различных величин коэффициента λ, а левая вертикальные точечная линия указывает минимальное значение MSE для выбранной модели. Правая вертикальная точечная линия показывает величину коэффициента λ, с минимальной дисперсией средней квадратичной ошибки.

Метод LASSO. Кроме *гребневого* метода в настоящее время широкое распространение получил метод *LASSO* (Least Absolute Selection and Shrinkage Operator), предложенный в работе (Tibshirani, 1996). Целевой функционал метода *LASSO* комбинирует среднеквадратичную ошибку линейной модели с суммой модулей всех коэффициентов:

$$\hat{\boldsymbol{\beta}}(\lambda) = \boldsymbol{arg}\left(\underset{\boldsymbol{\beta}}{\boldsymbol{min}}\left(\frac{\boldsymbol{1}}{\boldsymbol{2n}} \|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right) \right), \quad (3)$$

где $\|\cdot\|_1$ — l_1 норма вектора;

$\lambda > 0$ — регулирующий параметр, оптимальное значение $\hat{\lambda}$ которого может определяться по критерию среднеквадратичной ошибки:

$$\hat{\lambda} = arg(\underset{\lambda}{min}\left(\frac{\boldsymbol{1}}{\boldsymbol{2n}} \left\| \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda) \right\|_2^2 \right),$$

или как в *гребневом* методе с помощью кросс-валидации.

Метод *LASSO* не только стремится приблизить коэффициенты регрессии к нулю, но и может присваивать им нулевые значения. Таким образом, осуществляется выбор наиболее значимых предикторов. Рис. 2 показывает пример вычислительных путей коэффициентов и среднеквадратичной ошибки регрессии (MSE) в процессе подбора значения регулирующего параметра λ в методе *LASSO*.

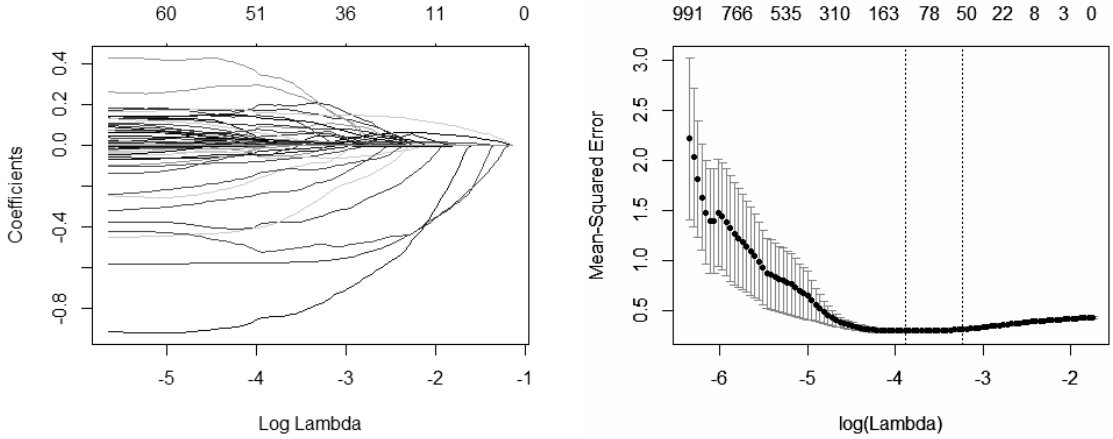

Рис. 2. Изменения величины коэффициентов регрессии и средней квадратичной ошибки (MSE) для метода *LASSO* в зависимости от значения регулирующего параметра λ. Левая точечная вертикальная линия на правом графике соответствует минимуму MSE и выбранной модели, правая — модели с наименьшей MSE вариацией.

Различие в поведении вычислительных путей метода *LASSO* и гребневого метода объясняется различиями в форме областей допустимых решений этих методов при фиксированном λ: выпуклый многогранник и эллипсоид (см. рис. 3, Hastie et al., 2015).

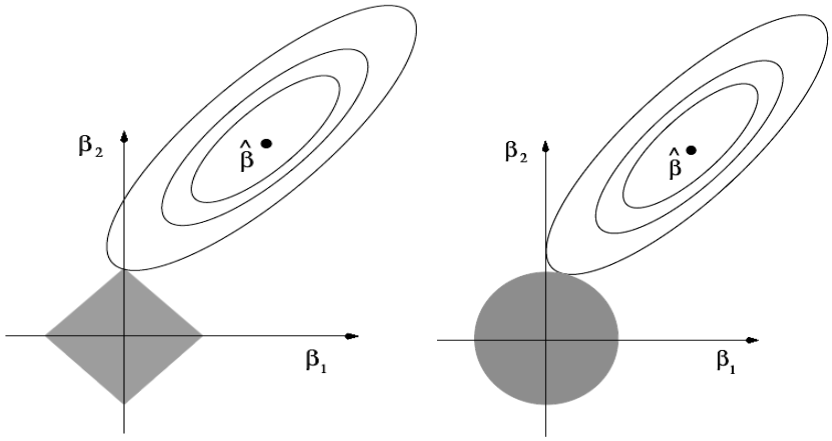

Рис. 3. Графическое представление условной задачи оптимизации для *LASSO* (слева) и ридж регрессии (справо). В случае *LASSO* регрессии допустимая область решения — область $|\beta_1| + |\beta_2| \le t$ с острыми углами. В случае ридж регрессии — шар $\beta_1^2 + \beta_2^2 \le t^2$. Точка $\hat{\beta}$ — обычная оценка методом наименьших квадратов.

Во многих случаях разреженность моделей *LASSO* рассматривается как преимущество, которое облегчает интерпретацию модели. В тоже время, *LASSO* не имеет аналитического решения, что делает более затруднительными вычислительные и теоретические результаты.

Модели LASSO часто дают меньшее значение среднеквадратичной ошибки прогноза (MSPE), чем модели *гребневого* метода. Но в случаях, когда $p > n$ метод *LASSO* отбирает не более n предикторов и игнорирует остальные независимо от их прогнозного потенциала. Поэтому *LASSO* может быть непригоден для извлечения информации из редких событий, если их количество превышает размер обучающей выборки. Кроме того, в *LASSO* отсутствует свойство группирования: отбирается только один предиктор из группы сильно коррелированных предикторов. Если $p > n$ и существуют группы коррелированных предикторов, то модели *LASSO* часто уступают гребневым моделям по прогнозной точности (MSPE).

Метод *Elastic Net*.. Zou, Hastie (2005) предложили объединить регулирующие добавки метода *LASSO* и *гребневого* метода путём их линейной комбинации. Новый метод получил название *Elastic Net* — эластичная сеть. В этом методе функционал решения имеет вид:

$$\hat{\boldsymbol{\beta}}(\lambda,\alpha) = arg\left(\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \left(\alpha \|\boldsymbol{\beta}\|_1 + \frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}\|_2^2 \right) \right) \right), \tag{4}$$

где $\|\cdot\|_2$ — l_2 норма вектора;

α — дополнительный регулирующий параметр $(0 \le \alpha \le 1)$, определяющий линейную выпуклую комбинацию гребневого метода и метода *LASSO*; при $\alpha = 0$ получаем гребневый метод, при $\alpha = 1$ — метод *LASSO*.

Оптимальная пара значений $(\hat{\alpha}, \hat{\lambda})$ определяется аналогично гребневому и *LASSO* методам по критерию среднеквадратичной ошибки:

$$(\hat{\alpha}, \hat{\lambda}) = arg\left(\min_{\alpha,\lambda} \left(\frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\alpha,\lambda)\|_2^2 \right) \right),$$

или с помощью кросс-валидации.

Elastic net преодолевает ограничения как *гребневого* так *LASSO* методов: обладает свойством группирования коррелированных предикторов и позволяет при $p > n$ включать в модель более n переменных.

Однако все вышеприведенные методы неспособны сами по себе включать в модель редкие значащие предикторы. Поэтому требуется введение в расчеты дополнительных параметров, позволяющих учитывать сходство между различными предикторами.

3. Кластеризация как основа агрегирования предикторов

Агрегирование предикторов по дополнительным признакам может позволить включить в регрессионную модель важные редкие предикторы. Важный редкий предиктор может попасть в регрессионную модель за счет суммирования значений всех предикторов, которые входят в один агрегированный предиктор. Такой агрегированный предиктор может уже не воспринимается как редкий и может получить ненулевое значение коэффициента в регрессионной модели.

Агрегирование может быть разного уровня, начиная с пустого, когда отсутствуют какие-либо объединения исходных предикторов, и заканчивая полным, при котором все исходные предикторы объединяются в один общий предиктор. Кроме уровня агрегирования может варьироваться и состав агрегированных групп. Критерием уровня и состава агрегирования выступает в конечном итоге тот же критерий, что и для регрессионной модели в целом. Обычно это среднеквадратичная ошибка прогноза MSPE. Значение MSPE при некотором агрегирования определяется как значение MSPE для регрессионной модели, построенной при этом агрегировании. Понятно, что такое значение зависит не только от принятого агрегирования, но и от метода построения регрессионной модели.

Важно, чтобы предикторы, входящие в один агрегированный предиктор, были достаточно «близкими» друг к другу по влиянию на отклик. Для целей агрегирования используется дополнительная информация о предикторах. В общем случае у предикторов может быть определенная совокупность характеристик, от которых зависит их влияние на отклик. Эти характеристики связаны со спецификой задачи и не всегда просто определяются. Если такие характеристики известны и оцифрованы, то появляется возможность размещения предикторов в пространстве их характеристик. В этом пространстве близкие (по расстоянию) предикторы должны «близко» влиять на отклик. Таким образом, агрегирование предикторов сводится к задаче определения групп предикторов, близких друг к другу в пространстве их характеристик, т.е. к задаче кластеризации предикторов.

Так как уровень и состав кластеризации подбираются в процессе построения регрессионной модели, то появляются дополнительные регулирующие параметры, а именно: метод (модель) кластеризации (иерархический, *k*-средних, спектральный и др.), критерий схожести (метрика пространства, другие характеристики), уровень (степень) кластеризации (пороговое значение метрики, число кластеров и др.). Обычно ограничиваются рассмотрением одного метода и, возможно, нескольких критериев схожести. При любом методе и критерии схожести универсальным показателем уровня кластеризации является число кластеров, по которому (при заданных методе и критерии схожести) однозначно определяются кластеры.

Обычно используется иерархическая агломеративная кластеризация (например, Grira et al., 2005), в которой строится дерево (дендрограмма) кластеризации. Корень дерева представляет кластер, включающий все предикторы, а листья дерева — отдельные предикторы. В качестве критерия объединения объектов кластеризации используется расстояние между ними в соответствующем многомерном пространстве. Конкретные кластеры определяются с помощью дополнительного параметра оптимизации: высота по дендрограмме или общее число кластеров. После этого строится регрессионная модель на агрегированных предикторах. Полученную модель пересчитывают в регрессионную модель на исходных предикторах, в которой коэффициент при каждом исходном предикторе принимается равным коэффициенту при агрегированном предикторе, содержащем этот исходный предиктор. Т.о., редкий

предиктор получает возможность попасть в регрессионную модель через агрегированный предиктор (например, Wang, Zhao, 2017).

На рис.4 приведен пример предварительного агрегирования группы английских слов на основе иерархической дендрограммы, построенной по 50-мерной модели английского языка GloVe (Pennington, 2014) обученной на данных Gigaword 5 и Wikipedia 2014 (GloVe, 2014). Горизонтальные линии дендрограммы показывают какие именно листья или ветви дендрограммы объединяются в один кластер. Высота дендрограммы (Height) отражает расстояние между объединяемыми словами и кластерами. Уровень кластеризации задается либо высотой по дендрограмме, либо общим количеством кластеров (подробнее см. раздел 5).

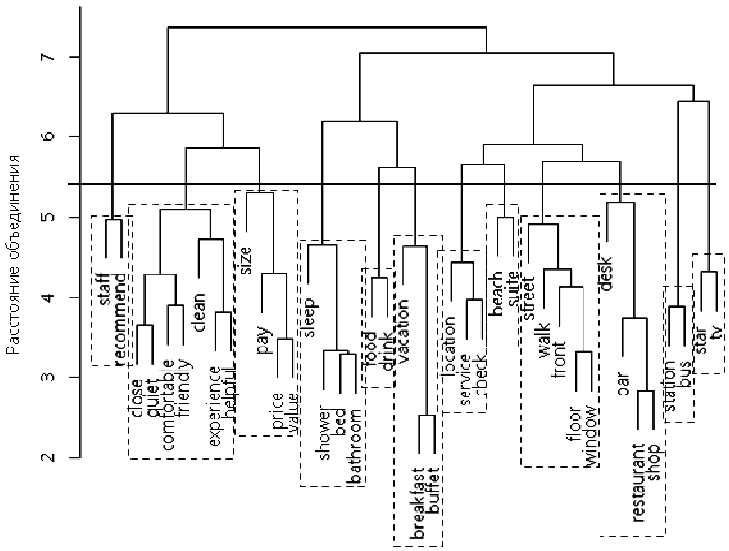

Рис. 4. Пример предварительного агрегирования предикторов на основе иерархической дендрограммы по высоте дендрограммы (расстояние объединения). Каждый кластер предикторов выделен пунктирным прямоугольником.

4. Методы регрессионного анализа с адаптивным агрегированием предикторов

При адаптивном агрегировании решение об объединении предикторов в один агрегированный предиктор принимается в процессе подбора регрессионной модели. Основой агрегирования является иерархическое дерево кластеризации исходных предикторов. Адаптивное агрегирование является неотъемлемой частью специальных методов подбора регрессионных моделей.

Метод регрессионного анализа с учетом дендрограммы кластеризации предикторов был предложен Wang, Zhao, 2017. Метод получил название *TASSO* (Tree-guided Automatic Subcomposition Selection Operator). В методе *TASSO* функционал решения имеет следующий вид:

$$\hat{\boldsymbol{\beta}}(\lambda_1,\lambda_2) = \boldsymbol{arg}\left(\min_{\boldsymbol{\beta}} \left(\frac{1}{2n} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1 + \lambda_2 \sum_{u \in \boldsymbol{U}} \left| f_u^T \cdot \boldsymbol{\beta} \right| \right) \right), \quad (5)$$

где $\boldsymbol{U}$ — множество внутренних вершин дерева кластеризации предикторов $\Im$,

$\boldsymbol{f}_u = \sum_{j \in \boldsymbol{L}_u} \boldsymbol{e}_j$ ($\boldsymbol{L}_u$ — множество листьев поддерева $\Im$ с корневой вершиной u; $\boldsymbol{e}_j$ — p-мерный вектор, у которого j-й элемент равен 1, а остальные равны 0 ($j = 1,...,p$),),

$\lambda_1 \geq 0$ и $\lambda_2 \geq 0$ — параметры регуляризации *TASSO* решения.

Оптимальную пару значений $(\hat{\lambda}_1, \hat{\lambda}_2)$ авторы метода *TASSO* рекомендуют определять с помощью одного из информационных критериев.

В формуле (5) первые два слагаемых совпадают со слагаемыми формулы *LASSO*. Третье слагаемое добавляет кратность значений $|\beta_j|$ в соответствии с деревом кластеризации $\Im$: каждое значение $|\beta_j|$ входит в третье слагаемое формулы (5) столько раз, сколько внутренних вершин дерева $\Im$ соединено с листом j (или, другими словами, во сколько разных кластеров может входит предиктор j в соответствии с дендрограммой). Т.о., метод *TASSO* стремиться занулить в первую очередь те $|\beta_j|$, j - предикторы которых входят в наибольшее число разных кластеров в соответствии с дендрограммой. Однако, это число не отражает возможную близость предиктором и его использование для группирования предикторов не обоснованное.

В целом, в методе *TASSO* не просматривается агрегирование предикторов на основе дерева кластеризации $\Im$, что не позволяет использовать этот метод для извлечения информации из редких событий.

Адаптивное агрегирование предикторов на основе иерархической кластеризации впервые предложено в методе *rare*, ориентированном на обоснованное включение редких предикторов в регрессионную модель (Yan, Bien, 2018).

Для описания этого метода рассмотрим общий вид дерева кластеризации предикторов $\Im$. У этого дерева p листьев и каждый лист соответствует определенному исходному предиктору. Каждому листу j $(j = 1,...,p)$ дерева ставится в соответствие коэффициент линейной регрессии β_j. Все коэффициенты листьев одной ветви, которая агрегируется, приравниваются между собой. Для этого вводится числовой параметр γ_u, который приписывается каждой внутреннему узлу u дерева $\Im$.

Значения γ_u и β_j связываются равенством

$$\beta_j = \sum_{u \in \{j\} \cup ancestor(j)} \gamma_u , \qquad (6)$$

где $\{j\} \cup ancestor(j)$ — множество вершин на пути от корня дерева к листу j ($ancestor(j)$ — множество вершин-предшественников листа j в дереве).

Т.е., коэффициенты β_j представляются в виде суммы γ_u вдоль пути $\{j\} \cup ancestor(j)$. И именно такое представление позволяет выйти на агрегирование предикторов.

Задача агрегирования предикторов рассматривается как задача разделения дерева $\Im$ на множество непересекающихся ветвей. Каждая ветвь в таком разделении представляет один агрегированный предиктор, в котором объединяются исходные предикторы-листья этой ветви. На рис. 5 показан пример агрегирования с учетом значений γ_u, приписанных внутренним вершинам.

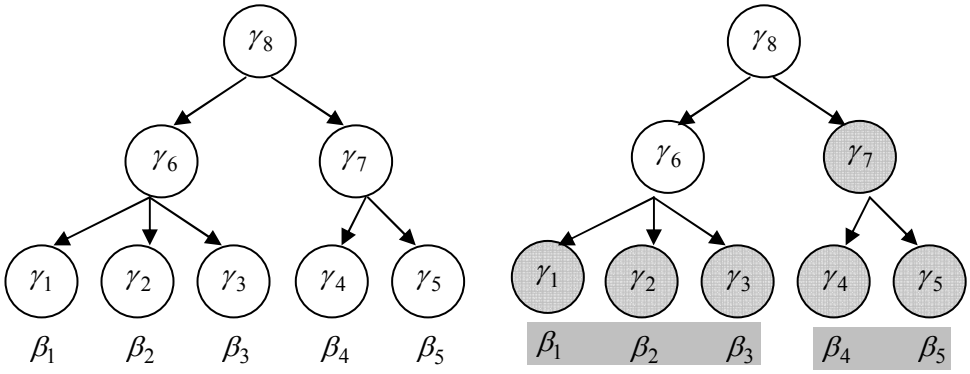

Рис. 5. Пример агрегирования предикторов на основе иерархического дерева в методе *rare*.

В этом примере пять исходных предикторов — это пять листьев дерева и $\beta_i = \gamma_i + \gamma_6 + \gamma_8$, $i = 1,2,3$; $\beta_j = \gamma_j + \gamma_7 + \gamma_8$, $j = 4,5$. При зануленении γ_j, соответствующих серым узлам дерева, исходные предикторы агрегируются в две группы: $\{1,2,3\}, \{4,5\}$ с коэффициентами $\beta_1 = \beta_2 = \beta_3 = \gamma_6 + \gamma_8$ и $\beta_4 = \beta_5 = \gamma_8$.

Равенство (6) может быть записано в матричной виде: $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma}$, где $\boldsymbol{\beta} = (\beta_1,...,\beta_p)^T$ — вектор-столбец коэффициентов; $\boldsymbol{\gamma} = (\gamma_{1,...,}\gamma_{|\Im|})^T$ — вектор-столбец параметров γ_u; $\boldsymbol{A} = \left(a_{jk}\right)_{i=1,...,p;\, k=1,...|\Im|}$ — бинарная матица, определяющая дерево $\Im$: $a_{jk} \in \{0,1\}$; $a_{jk} = 1$ для $k|u_k \in \{j\} \cup ancestor(j)$. Например, для дерева на рис.5 $\boldsymbol{A} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}$.

Вершины дерева, последующие при движении от корня к вершине u, определяют ветвь $\Im_u = \{u\} \cup descendant(u)$ ($v \in descendant(u)$ — множество вершин, последующих вершине u). При зануленении всех γ_v для $v \in descendant(u)$ коэффициенты β_j всех листьев ветви $\Im_u$ приравниваются между собой. Это равносильно агрегированию предикторов, соответствующих этим листьям.

Для определения оптимальных значений векторов $\boldsymbol{\beta}$ и $\boldsymbol{\gamma}$ решается следующая оптимизационная задача:

$$\hat{\boldsymbol{\beta}}(\lambda,\alpha) = arg\left(\min_{\boldsymbol{\beta}\in\boldsymbol{R}^p, \boldsymbol{\gamma}\in\boldsymbol{R}^{|\Im|}} \left(\frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\right\|_2^2 + \lambda\left(\alpha\left\|\boldsymbol{\gamma}_{-r}\right\|_1 + (1-\alpha)\left\|\boldsymbol{\beta}\right\|_1\right) \middle|\ \boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\gamma} \right) \right), \tag{7}$$

где n — размер выборки данных,

$\boldsymbol{Y}$ — вектор отклика ($\boldsymbol{Y} = (y_1,...,y_n)$),

$\boldsymbol{X} = \left(x_{ij}\right)_{i=1,...,n;\, j=1,...,p}$ — матрица значений предикторов (x_{ij} — значение предиктора j в i-ом испытании),

$\boldsymbol{\gamma}_{-r}$ — вектор параметров $\boldsymbol{\gamma}$ без параметра корневой вершины r,

$\alpha \in [0,1]$ и $\lambda > 0$ — параметры регуляризации.

В формуле (7) первое слагаемое представляет ошибку МНК прогноза. Второе слагаемое дает совместную регуляризацию значений $(\beta_1,...,\beta_p)$ и $(\gamma_{1,...,}\gamma_{|\Im|})$. При этом условие $\boldsymbol{\beta} = \boldsymbol{A\gamma}$ обеспечивает равенство коэффициентов β_j для предикторов из одного кластера. Количество и состав кластеров подбираются в процессе решения оптимизационной задачи (7) через значения $(\gamma_{1,...,}\gamma_{|\Im|})$. Такая постановка задачи позволяет с помощью l_1- регуляризации решения достигать разреженности ненулевых значений γ_u (т.е., агрегирование предикторов) и разреженности ненулевых значений β_j (т.е., отбор агрегированных предикторов).

Параметр регуляризации λ контролирует общий уровень штрафа, α — баланс между агрегированием и селекцией предикторов. Оба параметра подбираются в процессе кросс-валидации. При $\alpha = 0$ задача (7) сводится к задаче метода Lasso на $\boldsymbol{\beta}$, при $\alpha = 1$ — к обобщенной задаче метода Lasso на $\boldsymbol{\gamma}$ (Hastie et al, 2015)

Оптимальная пара значений $(\hat{\alpha}, \hat{\lambda})$ определяется аналогично гребневому методу, методам *LASSO* и *Elastic Net* с помощью внешней процедуры (кросс-валидации) по критерию среднеквадратичной ошибки:

$$(\hat{\alpha}, \hat{\lambda}) = arg(\min_{\alpha,\lambda}\left(\frac{1}{2n}\left\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}(\alpha,\lambda)\right\|_2^2\right)$$

Yan, Bien (2018) применили для решения задачи (7) метод множителей переменного направления (alternating direction method of multipliers — ADMM, Boyd et al., 2011). На основе этого метода был разработан R пакет программ *rare* (Yan, Bien, 2018a).

В теоретическом плане адаптивное агрегирование вышеуказанного метода выглядит более предпочтительным в сравнении с предварительным агрегированием, так как позволяет на разных ветвях дерева кластеризации выбирать разные уровни кластеризации. На рис. 6 дан пример возможного адаптивного агрегирования предикторов в методе *rare* для того же дерева кластеризации, что и на рис. 4 . В этом примере расстояние объединения кластеров варьируется по ветвям дерева. Из общих соображений адаптивное агрегирование расширяет область рассматриваемых решений, что позволяет надеяться на большую эффективность решений, чем при предварительном агрегировании.

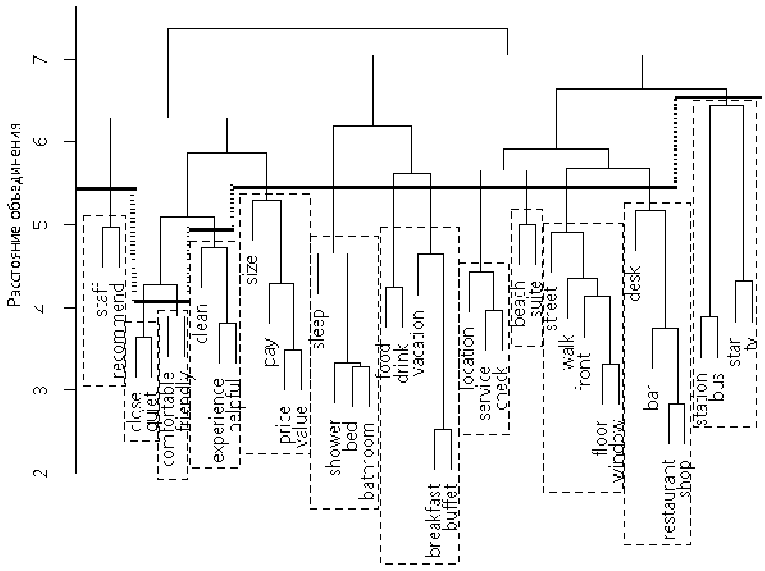

Рис. 6. Пример возможного адаптивного агрегирования предикторов на основе иерархической дендрограммы. Каждый кластер предикторов выделен пунктирным прямоугольником

Преимущества адаптивного агрегирования могут следовать из особенностей проявления дополнительных свойств предикторов, на основании которых строится дендрограмма, в конкретном регрессионном анализе (например, особенностей использования слов в определенной группе текстов).

Однако, адаптивное агрегирование предикторов в методе *rare* опирается только на топологию дендрограммы, не учитывая степень близости предикторов в том или ином кластере. Кластеры близких и не близких предикторов в методе *rare* равноценны. В теоретическом плане это может снижать эффективность этого метода.

5.Подготовка исходных данных

Основой исходных данных, по которым выполнялись расчеты, являются уже упомянутые реальные данные TripAdvisor Data Set (2009). Эти данные были подготовлены Wang et al. (2010) на основе информации с сайта TripAdviser.com. Данные включают в себя 235.793 отзыва (вместе с числовыми рейтингами от 1 до 5) о качестве проживания и обслуживания в 1850 отелях. Wang и др. обработали эти данные до состояния, пригодного к использованию в регрессионном анализе. В данных были оставлены только прилагательные — файл *Vector_shLDA_1999.txt.* В этом файле для каждого из 1850 отелей приведен общий усредненный рейтинг по всем отзывам и еще 7 усредненных рейтингов по разным аспектам проживания. Для каждого рейтинга дается матрица частот использования выделенных 1999 слов во всех отзывах.

Пример части содержания файла *Vector_shLDA_1999.txt,* относящейся к отелю hotel_1225375 (данные, относящиеся к одному отелю, разделены на 8 записей) переменной длины:

1-я запись (индекс отеля и 8 усредненных рейтингов)

hotel_1225375_parsed 3.46154 3.66667 3.66667 3.41667 3.5 3.75 3.41667 3.375

2-я запись (матрица частот слов в отзывах, относящихся к первому (общему) усредненному рейтингу)

3:2 5:1 7:4 9:4 12:1 13:1 15:1 18:1 22:2 25:2 26:1 35:1 39:1 44:3
49:1 53:3 59:2 70:3 76:2 86:3 104:1 128:3 138:2 148:2 159:1 170:2 186:1
196:3 203:1 229:1 232:1 234:2 248:1 249:2 261:1 267:1 287:1 293:1 310:1 321:3
332:2 464:1 465:1 478:1 510:1 563:1 598:1 638:1 639:2 642:1 656:1 721:1 738:1
797:1 811:1 874:1 881:1 895:1 906:1 937:1 941:1 984:1 1037:1 1112:1 1116:1
1125:1 1352:1 1358:1 1371:1 1379:1 1663:1 1664:1 1673:1 1740:1 1919:1

3-я запись (матрица частот слов в отзывах, относящихся ко второму усредненному рейтингу)

…

В этих данных число перед двоеточием — индекс слова в списке из файла *CVWords_1999.txt*, после двоеточия — количество употребления этого слова в отзывах по данному отелю.

Пример показывает, что в рассматриваемом отзыве общего рейтинга использовано всего 76 разных слов (прилагательных), из которых большая часть употреблена только один раз.

На основе данных из файла *Vector_shLDA_1999.txt* была подготовлена матрица частот терминов $\boldsymbol{X} = \left(x_{ij}\right)_{i=1,\ldots,1850;\, j=1,\ldots,1999}$, где x_{ij} — количество j-тых слов в общих отзывах по i-му отелю (2-я запись для i-го отеля), и вектор откликов $\boldsymbol{Y} = (y_1,\ldots, y_{1999})$, где y_i — усредненный общий рейтинг i-го отеля (первое число 1-ой записи для i-го отеля, $y_i \in [1;5]$).

Степень разреженности матрицы $X = \left(x_{ij}\right)_{i=1,\ldots,1850;\, j=1,\ldots,1999}$ отражается гистограммой на рис. 7.

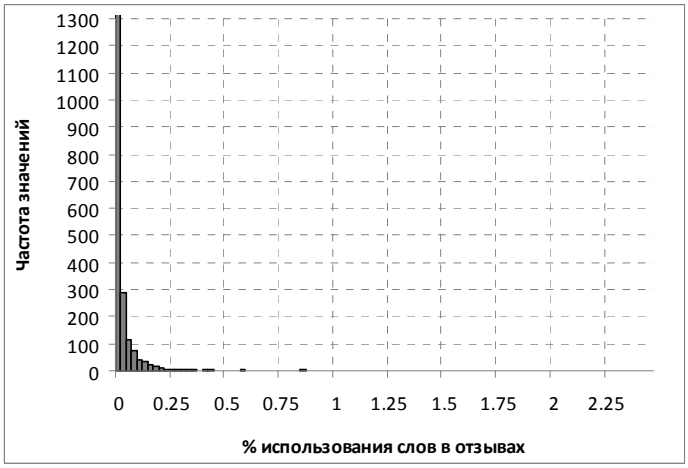

Рис. 7. Гистограмма частот использования слов в отзывах

Гистограмма показывает, что практически все термины из *CVWords_1999.txt* использованы в не более чем 0.25% обобщенных отзывов. Наиболее часто в отзывах появлялись такие термины: resort (3.59 %), price (2.73 %), experience (2.04 %), pay (1.93 %), star (1.42 %), money (1.38 %), … Т.о., влияние каждого отдельного термина на рейтинг ничтожно. Существенным может быть только объединенное влияние определенной совокупности терминов. Поэтому в подобных случаях важным является агрегирование терминов.

Так как $y_i \in [1;5]$, то дополнительно для матрицы X была применена простая линейная нормализация: $\bar{x}_{ij} = \dfrac{x_{ij}}{\sum\limits_{j=1}^{p} x_{ij}}$, где $x_{ij}, \bar{x}_{ij}$ — исходное и нормализованное значения предиктора; p — общее количество предикторов. В результате такой нормализации строки частот слов, заменены на строки структуры обобщенного отзыва по частоте использования слов.

Для оценки семантической близости слов из *CVWords_1999.txt* были использованы данные из GloVe (2014): файл *glove.6B.50d.txt* с 50-мерным векторным представлением 400000 слов. Небольшое количество слов из *CVWords_1999.txt* с ошибками в написании (порядка 50) отсутствовало в *glove.6B.50d.txt*. Такие слова были вручную скорректированы в *CVWords_1999.txt.*

Кроме того, были добавлены оценки эмоционального-чувственного восприятия английских слов из словаря *NRC-Emotion-Lexicon-Wordlevel-v0.92.txt* (NRC Emotion Lexicon, 2020), В этом словаре для каждого из почти 14200 слов отражены бинарные ассоциации с восьмью эмоциями (anger, fear, anticipation, trust, surprise, sadness, joy, disgust) и двумя чувствами (negative, positive): 1 — ассоциация есть, 0 — ассоциации нет. Ассоциации из *NRC-Emotion-Lexicon-Wordlevel-v0.92.txt* были включены в файл векторного представления слов (words_vector) путем увеличения его размерности на 7 единиц. Правила включения представлены в табл. 1.

Таблица 1. Правила включения эмоционально-чувственных ассоциаций в файл векторного представления слов

Компонент вектора слова	Векторизированное значения ассоциации		
words_vector [,51] =	–1 if negative=1	1 if positive=1	0 otherwise
words_vector [,52] =	–1 if sadness=1	1 if joy=1	0 otherwise
words_vector [,53] =	–1 if disgust=1	1 if trust=1	0 otherwise
words_vector [,54] =	–1 if fear=1		0 otherwise
words_vector [,55] =		1 if surprise=1	0 otherwise
words_vector [,56] =	–1 if anger=1		0 otherwise
words_vector [,57] =		1 if anticipation=1	0 otherwise

В первых трех строках эмоции и чувства разбиты на пары по принципу их противоположности и каждая пара отражается в файле *words_vector* одной размерностью. Остальные 4 эмоции считаются независимыми между собой. Численные значения (–1 или 1) эмоционально-чувственных ассоциаций в *words_vector* выбраны такими, чтобы быть существенными при кластеризации слов.

В результате мы получили объединенное векторное представление семантически-эмоционально-чувственной близости слов. На рис. 8 представлена гистограмма значений матрицы объединённого векторного представления семантически-эмоционально-чувственной близости слов.

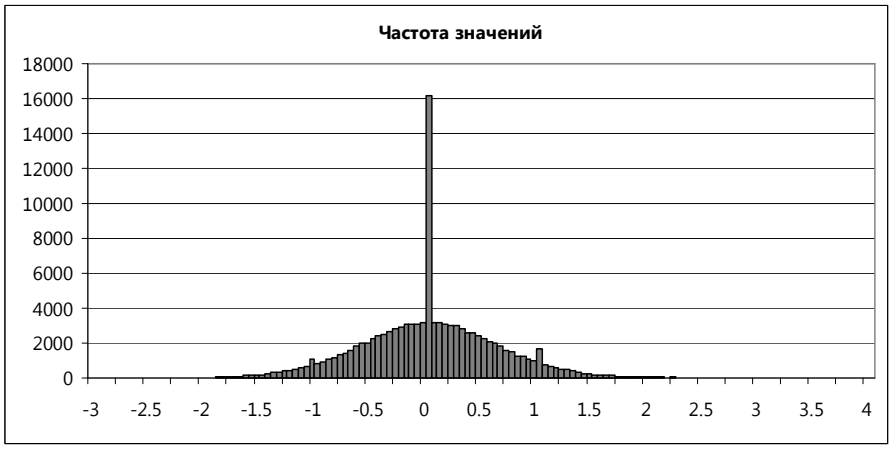

Рис. 8. Гистограмма значений матрицы векторного представления семантически-эмоционально-чувственной близости слов

Боковые и центральный пики на гистограмме рис. 8 вызваны дискретностью значений (–1;0;1) компонент *words_vector[,51:57].* Остальные значения по своему распределению близки к нормальному.

6.Результаты исследований

Авторами *rare* было выполнено вычислительное сравнение этого метода с методом *LASSO* в сочетании с предварительным агрегированием на реальных данных TripAdvisor Data Set (2009), описанных в предыдущем разделе.

Расчеты методом *LASSO* выполнялись как без предварительного агрегирования предикторов (обозначение *L1* в табл.1), так и с предварительным семантическим агрегированием предикторов на основе дерева кластеризации (обозначение *L1-ag*). При этом в методе *L1-ag* в сравнении с методом *L1* добавились еще два регуляризационных параметра: модель иерархической кластеризации и высота кластеризации по дендрограмме. Выделение кластеров делалось по высоте дендрограммы. Высота h дерева кластеризации, на которой выполнялось агрегирование данных, выбиралась из 10 равномерно распределенных точек интервала [0.001;0.1]. Сведения об использованном методе иерархической кластеризации не приведены. Обучающая выборка, по которой подбиралась регрессионная модель, изменялась от 1% до 100% полного набора данных. Подобранная модель тестировалась на полном наборе данных.

Результаты сравнения метода *rare* с методами *L1* и *L1-ag* приведены в табл. 2.

Таблица 2. Основные результаты сравнения применения метода *rare* с методами *L1* и *L1-ag*

Обучающее множество		Количество предикторов, p	n/p	Среднеквадратичная ошибка прогноза на полных данных (*mspe*)				
%	*N*			*rare*	*L1*	*L1-ag*	*rare* & *L1*	*rare* & *L1-ag*
1%	1.7	2.397	0.709	**0.870**	0.894	0.882	–2.68%	–1.36%
5%	8.499	3.962	2.145	**0.783**	0.79	0.785	–0.89%	–0.25%
10%	16.999	4.786	3.552	**0.758**	0.764	0.764	–0.79%	–0.79%
20%	33.997	5.621	6.048	**0.742**	0.749	0.747	–0.93%	–0.67%
40%	67.995	6.472	10.506	**0.739**	0.74	0.742	–0.14%	–0.40%
60%	101.992	6.962	14.650	**0.733**	0.736	0.734	–0.41%	–0.14%
80%	135.99	7.294	18.644	**0.733**	**0.733**	0.734	0.00%	–0.14%
100%	169.987	7.573	22.446	**0.729**	0.731	0.731	–0.27%	–0.27%

Таблица 2 показывает, что по критерию *mspe* (Mean Squared Prediction Error — среднеквадратичная ошибка прогноза) метод *rare* во всех случаях дал лучшие результаты, чем *L1* и *L1-ag*. Уменьшения ошибки прогноза *mspe* составило до 3%. Причем большее уменьшение значение *mspe* получается при меньшем размере обучающего множества — больше проявляется эффект агрегирования близких по влиянию предикторов.

Однако эти результаты не проясняют один из главных вопросов исследования: насколько информация из редких событий учитывалась в регрессионных моделях, какое количество редких предикторов включалось в регрессионные модели.

Осталось также неясным,

1) как ведет себя метод *rare* в сравнении с методом *EN-ag* (*Elastic Net* с предварительным агрегированием) ;
2) какие ограничения и недостатки есть у метода *rare*;
3) какие рекомендации по применению метода *rare* и др.

Кроме того, расчеты выполнялись при ненормированной матрице частот слов в отзывах X и нормированном векторе откликов Y ($y_i \in [1,...,5]$), что повышает ошибку регрессии *mspe* и может затенять результат сравнения методов.

Этих вопросы и неясности проясняются в независимых исследованиях, выполненные авторами данной статьи. Данные, использованные для расчетов, описаны в предыдущем разделе. Параметрами, определяющими результаты сравнения, выступают следующие две величины: n — размер обучающего множества, k — количество агрегированных предикторов (для методов с предварительным агрегированием).

Было выполнено сравнение методов *LASSO* (*L1-ag*), *Ridge* (*L2-ag*) и *Elastic Net* (*EN-ag*) с предварительным агрегированием предикторов и метода *rare*. Уровень агрегирования определялся заданием суммарного количества k кластеров (агрегированных предикторов) и оптимизированное значение находилось методом последовательных приближений в процессе расчетов. Модели, подобранные на одних и тех же обучающих множествах, проверялись на одних и тех же тестовых данных. Качество полученных моделей оценивалось по среднеквадратичной ошибке прогноза *mspe* на тестовых данных.

В предположении, что агрегирование будет полезнее при малых размерах обучающего множества, было задано исходное значение обучающего множества *n=185* (10% от размера полной выборки). При таком значении *n* были выполнены расчеты по подбору значения *k* методами *L1-ag* (r-функция Lasso()), *L2-ag* (r-функция cv.glmnet(…, alpha=0)) и *EN-ag* (r-функция cva.glmnet(…, alpha = seq(0, 1, len = 11)). Лучший результат, как ожидалось, получен при использовании метода *EN-ag* при *k=800*. Результаты метода *L1-ag* были незначительно хуже на большей части диапазона изменения уровня кластеризации. Метод *L2-ag* дал существенно худшие результаты. Результаты расчетов приведены на рис. 9.

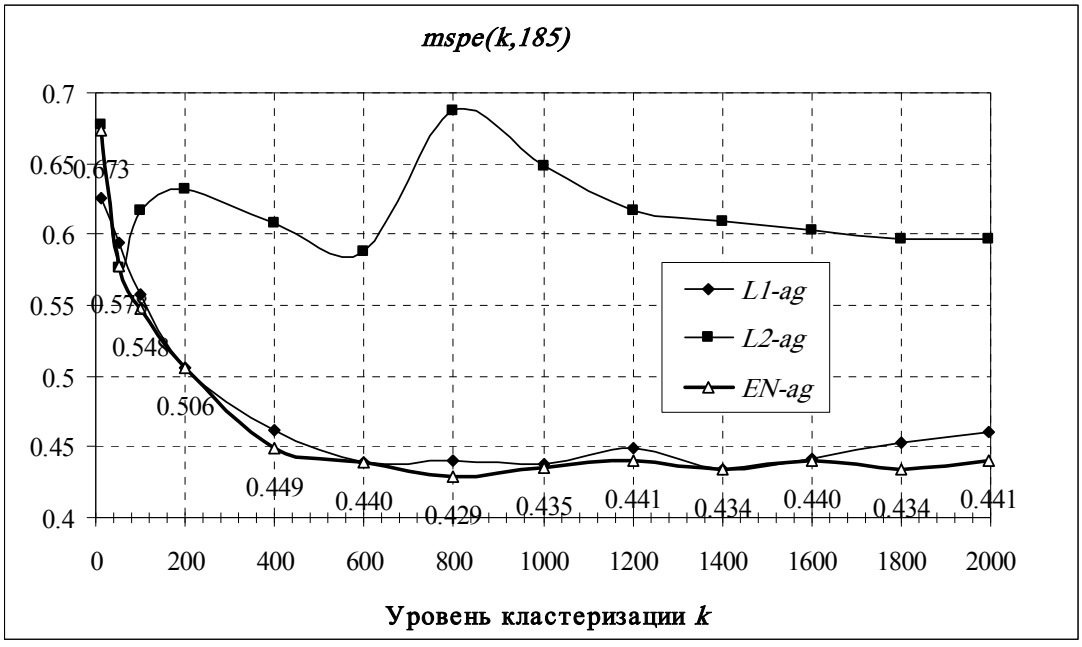

Рис. 9. Графики зависимости ошибки прогноза *mspe(k,185)* от уровня кластеризации на полном тестовом множестве. При *k=1999* агрегирование отсутствует, при *k= 1* — один агрегированный предиктор.

График на рис. 9 показывает изменение среднеквадратичной ошибки прогноза в зависимости от уровня агрегирования. Влияние уровня агрегирования предикторов на среднеквадратичную ошибку прогноза *mspe* оценивалось по относительному изменению ошибки (проценту изменения) в сравнении с отсутствием агрегирования: $\delta(mspe(k,n)) = \dfrac{mspe(k,n) - mspe(n)}{mspe(n)}$, где $mspe(k,n)$ — ошибка *mspe* на полном тестовом множестве при уровне агрегирования k и размере обучающего множества n; $mspe(n)$ — ошибка *mspe* на полном тестовом множестве при отсутствии агрегирования и размере обучающего множества n. Графики относительного изменение ошибки прогноза $\delta(mspe(k,185))$ приведены на рис. 10.

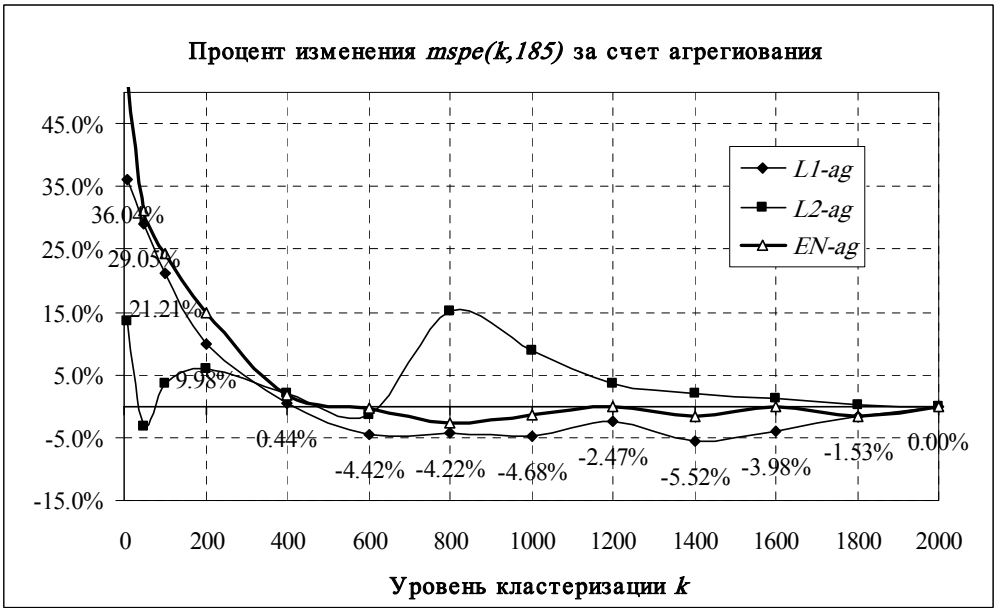

Рис. 10. Графики относительного изменения ошибки *mspe(k,185)* за счет агрегирования предикторов.

Положительный эффект агрегирования наблюдался у методов *L1-ag, EN-ag* при $k \geq 600$. Метод *L1-ag* оказался самым отзывчивым на агрегирование. При *k=1400* уменьшение в сравнении с отсутствием агрегирование составило более 5.5%.

На следующем шаге на основании результатов расчетов для *EN-ag* на обучающем множестве размера *n=185* был уточнен размер обучающего множества n по критерию $\delta(mspe(k,n)) = \left.\dfrac{mspe(k,n) - mspe(n)}{mspe(n)}\right|_{k=800} \rightarrow min$. Влияние размера обучающего множества на качество модели *EN-ag* при *k=800* приведено на рис. 11

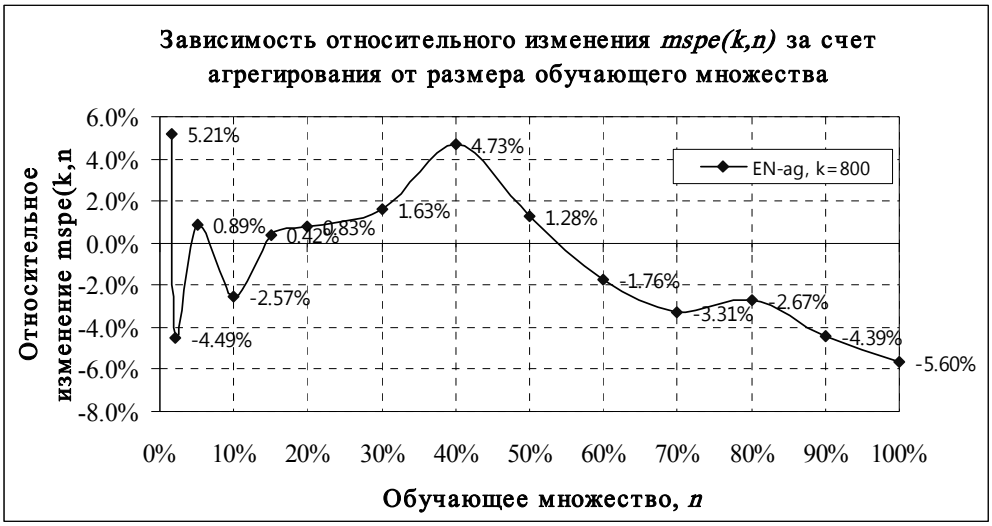

Рис. 11. Влияние размера обучающего множества на качество модели *EN-ag* при *k=800*

Рисунок показывает, что агрегирование дает устойчивый положительный эффект начиная с 55% размера обучающего множества. Естественно, что наибольшее снижение *mspe* (около 6%) получается при 100%-м размере обучающего множества.

Таким образом, предварительное семантическое агрегирование в нашем случае оказалось наиболее полезным в случае 100% обучающего множества. Максимальное уменьшение *mspe* за счет агрегирования при *k=1650* составило 17.63%. Влияние уровня кластеризации на качество модели при *n=1850* приведены на рис. 12.

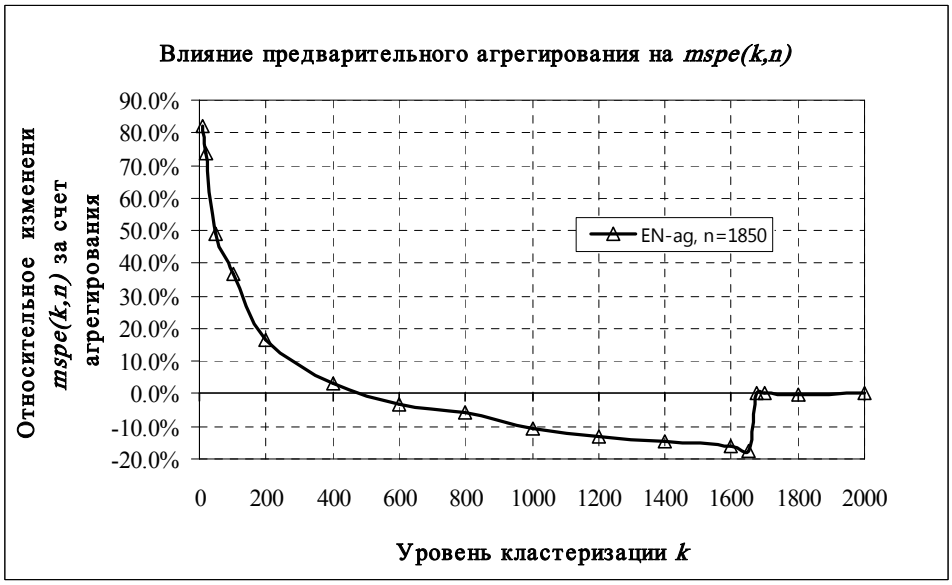

Рис. 12. Влияние уровня кластеризации на качество модели при *n=1850*

На основании проведенных расчетов метод *EN-ag* был выбран базовым для сравнения с методом регрессионного анализа, использующим адаптивное агрегирование, — методом *rare*.

Для расчетов методом *rare* использовались функции пакета ‘rare’ (Yan, Bien, 2018a). Основной была функция rarefit(y, X, hc, alpha = seq(0, 1, len = 11), rho=0.01,eps1 = 0.01,eps2 = 0.01, maxite = 10), где y — вектор отклика, X — матрица значений исходных предикторов, hc — результат построения иерархического дерева кластеризации предикторов функцией hclust(), alpha = seq(0, 1, len = 11) — последовательность задаваемых значений α , rho,eps1,eps2 — параметры регулирующие точность расчетов, maxite — количество итераций в расчетах.

Графики на рис. 13. наглядно сравнивают качество решений, полученных с помощью метода *rare* и некоторых вариантов метода *EN-ag*. В целом подавляющего преимущества метода *rare* не обнаруживается. Только на половине расчетных точек у метода *rare* меньшие значения *mspe*. В абсолютном сравнении (при 100%-м обучающем множестве) метод *rare* (при alpha = seq(0, 1, len = 11)) оказался только на третьем месте после *EN-ag* при *k=1650* и *k=1400*..

Кроме сравнения методов по *mspe* важно также сравнение по количеству предикторов, включенных в модель (отражается степень извлечения информации из редких предикторов). Информацию к этому вопросу дает табл. 3.

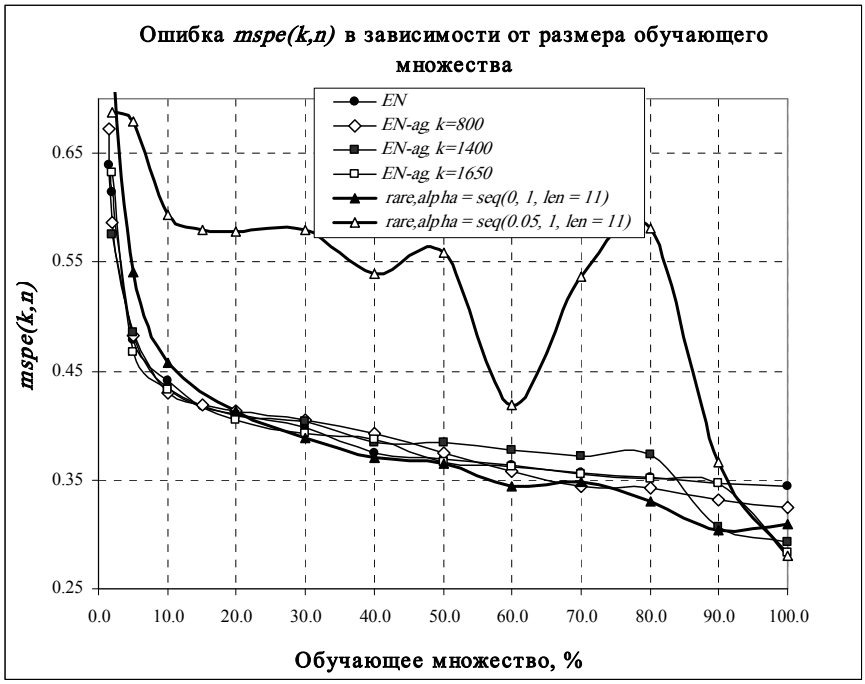

Рис. 13. Графики зависимости ошибки *mspe* от размера обучающего множества для основных вариантов расчетов

Таблица 3. Общие параметры основных построенных моделей (p_m , $p_{a,m}$ — количество исходных и агрегированных предикторов соответственно, включенных в модель)

n,%	*n*	*EN*			*EN-ag, k=1650*				*rare*			
		p_m	*mspe*	α	$p_{a,m}$	p_m	*mspe*	α	$p_{a,m}$	p_m	*mspe*	α
2	37	333	**0.6136046**	**0.5**	285	454	0.6327947	0.5	1939	**1939**	0.7293755	0.4
5	93	**234**	0.4783918	0.8	**202**	**334**	**0.4679983**	**0.8**	27	27	0.541142	0
10	185	170	0.440773	0.9	155	**259**	**0.4326426**	**0.9**	111	111	0.457456	0
20	370	204	0.4103834	0.9	181	**291**	**0.4049784**	**0.9**	113	113	0.4127288	0
30	555	193	0.3984722	0.9	169	**265**	0.3929369	0.9	135	135	**0.388175**	0
40	740	146	0.3745239	1	161	**265**	0.3867631	0.9	174	174	**0.3702802**	0
50	925	139	0.3692508	1	129	**216**	0.365625	1	159	159	**0.3646065**	0
60	1110	145	0.3640298	1	131	**225**	0.3619919	1	215	215	**0.3441805**	0
70	1295	140	0.3564798	1	126	**217**	0.3556479	1	162	162	**0.3486469**	0
80	1480	140	0.3526159	1	124	208	0.3516977	1	225	**225**	**0.3298097**	0
90	1665	152	0.3473941	1	133	234	0.3467915	1	294	**294**	**0.3046318**	0
100	1850	149	0.3435236	1	**1176**	**1491**	**0.2829503**	**0.4**	263	263	0.3099592	0

В таблице жирным текстом выделены лучшие результаты по строке. Параметр α для метода *rare* берется из формулы (7), для метода *Elastic Net* (*EN-ag*) — из функционала этого метода

$$\left(\frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda\left(\alpha\|\beta\|_1 + (1-\alpha)\|\beta\|_2^2\right) \right).$$

По данным табл. 3 метод *rare* был немного лучше метода *EN-ag* по критерию *mspe*. Однако, уступил по количеству предикторов, включенных в модель. Более того, метод *rare* совсем не использовал иерархическое дерево кластеризации ($p_{a,m} = p_m$ и $\alpha = 0$) — не объединял близкие предикторы. Таким образом, он отказался решать исходную задачу извлечения информации из редких событий — задачу, ради решения которой этот метод был создан.

Мы попробовали помочь методу *rare* и провели серию расчетов с исключением нулевого значения параметра регуляризации α : alpha = seq(0.05, 1, len = 11) (рис. 13). В итоге мы ушли от простого метода *LASSO*, но в результате существенно увеличилась ошибка *mspe* и обнаружились другие неприятности (табл. 4)

Таблица 4. Основные характеристики результатов расчетов методом *rare* при alpha = seq(0.05, 1, len = 11)

n,%	*n*	$p_{a,m}$	p_m	*mspe*	α	*Параметр регуляризации* λ	β_0	*Наиболее влияющие предикторы и значения их коэффициентов* β
2	37	1928	1928	0.6872541	0.525	5.238385e–07	3.553754	park, 1.488184 pay, –2.594992
5	93	**216**	**1999**	0.6799184	1	2.701407e–06	6.650322e+13	eat, –6.650322e+13 buck, –6.650322e+13
10	185	1998	1998	0.5930696	0.905	4.703828e–07	3.883131	value, 1.334948 price, –2.518262
15	278	1993	1993	0.5791362	0.62	4.156766e–07	3.91413	worth, 1.156887 price, –2.196177
20	370	1990	1990	0.5786861	0.62	3.92141e–07	3.922492	paris, 1.439647 price, –1.853137
30	555	1999	1999	0.5801695	0.905	3.690174e–07	3.924183	paris, 1.404526 money, –1.754345
40	740	**399**	**1999**	0.5397609	1	7.930203e–06	1.908111e+14	parrot, –1.908111e+14 ruin, –1.908111e+14
50	925	1935	1935	0.5581646	0.05	4.253159e–07	3.959853	paris, 1.972399 bad, –2.380247
60	1110	**387**	**1999**	0.4193606	1	1.485415e–05	1.179135e+14	advisor, –1.179135e+14 awful, –1.179135e+14
70	1295	1941	1941	0.5368197	0.05	4.182736e–07	4.081288	excellent, 2.445177 bad, –3.190706
80	1480	**324**	**1999**	0.5812059	1	1.246698e–05	2.165938e+14	west, –2.165938e+14 horrible, –2.165938e+14
90	1665	**113**	**1999**	0.3669932	1	9.723971e–05	2.256358e+13	wonderful, –2.256358e+13 ruin, –2.256358e+13
100	1850	**583**	**1999**	0.2805076	1	5.935045e–06	1.038103e+14	forbid, –1.038103e+14 terrible, –1.038103e+14

В таблице жирным текстом выделены агрегированные предикторы.

Агрегирование предикторов происходило только при $\alpha = 1$. Однако, при этом получались огромные значения коэффициентов регрессионной модели: e+13, e+14. В таблице значения такие коэффициентов одного расчета равны между собой, но они различаются знаками меньших разрядов.

7. Выводы

В работе проведен сравнительный анализ четырёх основных методов обоснованного включения редких предикторов в регрессионную модель: метод *LASSO* с предварительным агрегированием, метод Тихонова (*гребневый* метод) с предварительным агрегированием, метод *Elastic Net* с предварительным агрегированием и метод *rare* с адаптивным агрегированием.

На основании проведенных расчётов метод *Elastic Net (EN-Ag)* показал наилучшие результаты по критерию минимальной среднеквадратичной ошибки прогноза и количеству редких предикторов включённых в модель. В этом методе для включения редких предикторов в регрессионную модель использовалось агрегирование на основе близости между редкими предикторами по дополнительным параметрам а иерархическом дереве кластеризации предикторов, а именно расстоянию между векторами слов по данным Glove и на основании семантических оттенков слов. Агрегирование «близких» между собой редких предикторов позволило существенно уменьшить среднеквадратичную ошибку прогноза (до 20% на рассмотренных данных) линейной регрессионной модели.

Несмотря на то, что в теоретическом плане адаптивное агрегирование может быть более качественным, чем предварительное, программная реализация метода *rare* в функциях rarefit(), rarefit.cv() в целом не показала убедительного преимущества перед методом *Elastic Net* с предварительным агрегированием. Однако метод *rare* в функциях rarefit(), rarefit.cv() имеет следующие два важных преимущества: автоматический выбор уровня агрегирования предикторов по критерию среднеквадратичной ошибки и, похоже, более удачную реализацию метода *LASSO*, в который метод *rare* вырождается при $\alpha = 0$. Существенными недостатками программной реализации метода *rare* в функциях rarefit(), rarefit.cv() являются:

- повышенные требования (в сравнении с функциями glmnet(), cva.glmnet() и др. метода *LASSO*) к вычислительным возможностям компьютера: памяти требуется на порядок больше и на два порядка увеличивается время вычислений;
- при обычном диапазоне допустимых значений $\alpha \in [0;1]$ выбирается обычно $\alpha = 0$ (метод *rare* работает как метод *LASSO*);
- при $\alpha \in [0;1)$ отсутствует агрегирование предикторов и существенно увеличивается ошибка *mspe*;
- при $\alpha = 1$ исходные предикторы агрегируются; однако, ошибка *mspe* увеличивается в сравнении с $\alpha = 0$ и значение коэффициентов подобранной модели становятся нереально большими: e+13, e+14.

В целом авторы считают, что для дальнейших исследований, связанных с включением редких предикторов в регрессионную модель, разумно отдавать предпочтение методу *Elastic Net* (*EN-ag*) с предварительным агрегированием предикторов

Благодарности. Авторы выражают благодарность профессору Маастрихтского университета Stephan Smeekes за помощь в выборе темы исследований и полезные консультации.

Список литературы

Тихонов А. Н. (1963). О регуляризации некорректно поставленных задач, *Докл. АН СССР*, 1963, том 153, №1, 49–52.

Alan Talevi et al. (2020). Machine Learning in Drug Discovery and Development. Part 1: A Primer. CPT: *Pharmacometrics & Systems Pharmacology*, 9, 129–142. On line at: https://ascpt.onlinelibrary.wiley.com/doi/pdf/10.1002/psp4.12491.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122.

GloVe: Global Vectors for Word Representation. (2014). On line at: https://nlp.stanford.edu/projects/glove/ .

Feinerer, I. and Hornik, K. (2017). tm: Text Mining Package. R package version 0.7–1.

Hastie, T., Tibshirani, R.J. and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 362 p. On line at: https://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS_corrected_1.4.16.pdf.

Hoerl, A. and Kennard, R. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12, 55–67. On line at: https://doi.org/10.1080/00401706.1970.10488634.

Hui Zou and Trevor Hastie, Regularization and variable selection via the elastic net, *J. R. Statist. Soc. B* (2005) 67, Part 2, pp. 301–320. On line at: https://hastie.su.domains/Papers/B67.2%20(2005)%20301–320%20Zou%20&%20Hastie.pdf.

NRC Emotion Lexicon. (2020) On line at: https://archive.org/details/nrc-emotion-lexicon-v0.92.

Ingo Feinerer et al. (2023) Package 'tm'. On line at: https://cran.r-project.org/web/packages/tm/tm.pdf.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. On line at: https://nlp.stanford.edu/pubs/glove.pdf.

Tibshirani, R.J. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Series B, 58: 267–288.

TripAdvisor Data Set. (2009). On line at: http://times.cs.uiuc.edu/~wang296/Data/.

Wang, H., Lu, Y., and Zhai, C. (2010). Latent aspect rating analysis on review text data: A rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10*, 783–792, New York, NY, USA. ACM.

Wang, T.E., and Zhao, H. (2017). Structured subcomposition selection in regression and its application to microbiome data analysis. *Ann. Appl. Stat.*, 11(2):771–791. On line at: https://www.researchgate.net/publication/312910482_Structured_subcomposition_selection_in_regression_and_its_application_to_microbiome_data_analysis.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., and Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature genetics*, *45*(10), 1113–1120. doi:10.1038/ng.2764. On line at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3919969/.

Yan, X. and Bien, J. (2018). Rare Feature Selection in High Dimensions. ArXiv e-print 1803.06675. On line at: https://arxiv.org/abs/1803.06675.

Yan, X. and Bien, J. (2018a). Package 'rare'. On line at: https://cran.r-project.org/web/packages/rare/rare.pdf.

Yang, Y. and Pederson, J.O. (1997). A Comparative Study on Feature Selection in Text Categorization. *Proceedings of 14th International Conference on Machine Learning, Nashville, 8–12 July 1997*, 412–420. On line at: http://courses.ischool.berkeley.edu/i256/f06/papers/yang97comparative.pdf.

Dushyn O., Dushyn B. **Extracting information from rare events in regression analysis**.

Dushyn Oleksiy – analytical worker at xc-elitesports.com, info@xc-elitesports.com

Dushyn Borys – member of House of scientists, Kharkov, Ukraine; boris_dushin@yahoo.co.uk

Abstract

This paper investigated an important practical problem of extracting information from rare events in sparse and high-dimensional data while building a linear regression model. It analyzes the advantages and the limitations of the different linear regression method used for high-dimensional problems. Main known methods were selected and tested on the real Tripadvisor.com dataset. The results of this research show the importance of the data aggregation based on hierarchical clustering. It allows extracting information from rare features by aggregating them according the clustering tree. Comparative analyses of main different linear regression methods that use clustering aggregation were done.

Keywords: rare events, regression analysis, sparse data, high-dimensional data, Lasso, Ridge, ElasticNet, rare methods, text mining, semantic aggregation, hierarchical clustering, vector word representation.

Jel classification: C51, C63, C87.