DOI: 10.15593/2499-9873/2022.1.09
УДК 519.173

А.Н. Кисляков

Российская академия народного хозяйства и государственной
службы при Президенте Российской Федерации,
Владимирский филиал, Владимир, Россия

ОТБОР ПРИЗНАКОВ ДЛЯ ИСПОЛЬЗОВАНИЯ В МОДЕЛЯХ ПРЕДИКТИВНОЙ АНАЛИТИКИ ВНЕШНЕЭКОНОМИЧЕСКОЙ ДЕЯТЕЛЬНОСТИ РЕГИОНОВ

Рассматривается актуальная проблема конструирования и отбора признаков в задачах построения прогностических моделей показателей внешнеэкономической деятельности регионов. Целью работы является разработка методики применения графовых моделей и методов понижения размерности для отбора признаков при построении моделей предиктивной аналитики внешнеэкономической деятельности регионов

В качестве подхода, позволяющего описать структуру внешнеэкономических связей, использовались графовые модели, реализующие возможность построения на основе алгоритма уменьшения размерности UMAP. Для построения оптимального графа на основе алгоритма UMAP необходимо варьировать количество ближайших соседей для каждой вершины и минимальное метрическое расстояние для установления связи между вершинами. Показано, что кликовый коэффициент симметрии графа позволяет оценить локальную связанность точек в построенном графе, формируя обобщенное представление о структуре сети с позиции наличия в ней кластеров. Индекс Джини графа позволяет дать оценку соответствия глобальной структуры графа реальным сетям. Отбор признаков осуществляется на основе анализа клик графа, обеспечивающих максимум взаимной информации *MI* при минимуме признаков, что максимально уменьшает искажения при описании структуры региональных внешнеэкономических связей.

Применение описанного подхода позволило устранить мультиколлинеарность признаков, осуществить отбор показателей, расширить возможности использования имеющегося набора данных за счет включения новых показателей, вносящих в модель новую полезную информацию о предметной области. Рассмотренный в работе метод отбора признаков рационально использовать для построения интерпретируемых прогностических моделей показателей ВЭД и как один из способов понижения размерности пространства признаков модели.

Полученные результаты позволяют сделать вывод о преимуществах рассмотренного подхода к реализации отбора признаков при построении прогностических моделей показателей внешнеэкономической деятельности регионов.

Ключевые слова: теория графов, отбор признаков, прогностические модели, понижение размерности, мультиколлинеарность, взаимная информация, коэффициент Джини, внешнеэкономические связи регионов, взаимная информация, кластеризация.

A.N. Kislyakov

Russian Academy of National Economy and Public Administration
under the President of the Russian Federation. Vladimir branch,
Vladimir, Russian Federation

FEATURES SELECTION FOR USE IN PREDICTIVE ANALYTICS MODELS OF REGIONS FOREIGN ECONOMIC ACTIVITY

The work is devoted to the actual problem of designing and selecting features in the tasks of constructing predictive models of indicators of foreign economic activity of regions. The aim of the work is to develop a methodology for the use of graph models and dimensionality reduction methods for the selection of features in the construction of predictive analytics models of foreign economic activity of regions.

As an approach to describe the structure of foreign economic relations, graph models were used that implement the possibility of building on the basis of the UMAP dimension reduction algorithm. To build an optimal graph based on the UMAP algorithm, it is necessary to vary the number of nearest neighbors for each vertex and the minimum metric distance to establish a connection between the vertices. It is shown that the clique symmetry coefficient of the graph makes it possible to estimate the local connectivity of points in the constructed graph, forming a generalized idea of the network structure from the position of the presence of clusters in it. The Gini index of the graph allows us to assess the correspondence of the global structure of the graph to real networks. The selection of features is based on the analysis of the graph clicks, which provides maximum mutual information *MI* with a minimum of features, which maximally reduces distortions in describing the structure of regional foreign economic relations.

The application of the described approach made it possible to eliminate the multicollinearity of features, to select indicators, to expand the possibilities of using the existing data set by including new indicators that introduce new useful information about the subject area into the model. The method of feature selection considered in the paper can be rationally used to construct interpreted predictive models of foreign economic activity indicators and as one of the ways to reduce the dimension of the model feature space.

The results obtained allow us to conclude about the advantages of the considered approach to the implementation of the selection of features in the construction of predictive models of indicators of foreign economic activity of regions.

Keywords: graph theory, feature selection, predictive models, dimension reduction, multicollinearity, mutual information, Gini coefficient, foreign economic relations of regions, mutual information, clustering.

Введение

Анализ внешнеэкономической деятельности регионов неразрывно связан с задачами планирования и прогнозирования комплекса экономических показателей и построением систем поддержки принятия решений и развития региональных внешнеэкономических связей на основе математических и инструментальных методов моделирования.

Применение методов интеллектуального анализа данных и моделей машинного обучения в задачах построения прогностических моделей показателей внешнеэкономической деятельности регионов, в

первую очередь, требует отбора признаков международных транзакций и изучения структуры региональных внешнеэкономических связей в целях устранения высокой степени линейной зависимости между переменными модели – мультиколлинеарности [1; 2]. Данное свойство негативно виляет на качество обучения модели, внося дополнительные «шумы», приводящие к неустойчивости коэффициентов прогностической модели.

Данное явление не столь критично для моделей глубокого обучения, однако для анализа тенденций изменчивости структуры региональных внешнеэкономических связей [3; 4] и оценки траектории развития региона относительно выпуска конкретных видов продукции и реализации экспортного потенциала региона необходимо применение интерпретируемых моделей для адекватной оценки влияния каждого признака транзакций на результат.

Специфика изучения структуры внешнеэкономических связей регионов предполагает использование ряда показателей: как первичных, основанных на объемах импорта и экспорта товаров по странам, товарным группам, так и вторичных агрегированных показателей, характеризующих сложность экспортной корзины. Указанная проблема требует развития существующих подходов к отбору признаков и обусловливает актуальность исследований в данном направлении.

Целью работы является разработка методики применения графовых моделей и методов понижения размерности для отбора признаков при построении моделей предиктивной аналитики внешнеэкономической деятельности регионов.

Теория

Традиционно для выявления мультиколлинеарности признаков используются подходы, основанные на методах классической статистики, а именно на расчете коэффициента парной корреляции Пирсона r по всем признакам набора данных [5]. В реальных практических задачах порог абсолютного значения коэффициента корреляции, при котором два признака являются мультиколлинеарными, может варьироваться в зависимости от условий, что несет в себе дополнительные трудности при отборе признаков.

Еще одним статистическим подходом для сокращения количества признаков является VIF-анализ (variance inflation factor) [1; 6], в основе которого лежит дисперсионный анализ: если при включении в модель новых факторов переменные являются ортогональными в признаковом пространстве, то коэффициент детерминации и переменные не снижают значимость друг друга. Несмотря на то что коэффициент множественной детерминации позволяет определить минимально необходимый набор признаков для обучения модели с высокой точностью, это не означает, что часть важных факторов, оказывающих влияние на качество предсказания модели на тестовой выборке, не будет упущена.

В случае, если имеется высокая нелинейная зависимость между факторами, на помощь приходят методы понижения размерности [1], реализующие переход от множества факторов к некоторым комбинациям признаков. Класс алгоритмов понижения размерности является одним из ключевых направлений развития классических методов машинного обучения на протяжении многих лет, начиная с анализа главных компонент (principal component analysis PCA) набора данных и заканчивая одним из наиболее продвинутых методов стохастического вложения соседей с *t*-распределением (t-distributed Stochastic Neighbor Embedding – t-SNE) [1].

Одним из подходов к формированию признаков является анализ графов, позволяющих описывать сетевую структуру региональных внешнеэкономических отношений по странам, группам товаров товарной номенклатуры внешнеэкономической деятельности (ТНВЭД) и формирующих сообщества (кластеры) [7]. Вершинами графа могут выступать как отдельные страны – экспортеры / импортеры продуктов, либо группы товаров, а ребрами – например, объем импортируемых / экспортируемых продуктов в денежном выражении и другие числовые и нечисловые показатели, характеризующие транзакции.

Данные и методы

Таможенная статистика по внешнеторговым отношениям регионов России собирается Федеральной таможенной службой в соответствии с утвержденной методологией для членов Евразийского экономического союза [8; 9]. Сформированный набор данных является основным источником информации о внешнеэкономической деятель-

ности Российской Федерации в разрезе федеральных округов и регионов и служит основой для анализа социально-экономического развития регионов России, стратегического планирования и прогнозирования. Набор формируется в виде таблицы, содержащей следующие поля: Направление: импорт / экспорт, период, страна-импортер / экспортер, товарные группы ТНВЭД, единицы измерения, стоимость, долл. США, масса нетто, кг, количество (если измеряется в шт.), регион и федеральный округ. Очевидно, что для объективной оценки объемов импорта / экспорта по отдельному региону рационально использовать показатель стоимости по отдельным периодам (месяцам), для этого необходимо построить сводную таблицу по товарным группам.

Основная проблема при построении графа, отображающего структуры внешнеэкономических связей региона, состоит в огромном количестве вершин: когда много товаров и потребителей, возникают сложности как при оценке большого количества связей, так и учета относительно «слабых» взаимодействий в сетевой модели. При большом количестве вершин V и ребер E анализ графа $G(V, E)$ усложняется, и его геометрическое представление не всегда удобно для анализа. Задача оптимизации структуры графа с позиции его информативности идентична задаче отбора признаков.

Следует также отметить, что описание взаимодействий по международным транзакциям на основе данных таможенной статистики возможно путем формирования групп временных рядов по странам, группам ТНВЭД, регионам. Эти три вида объектов формируют набор временных рядов по признакам в виде сводной таблицы, описывающих внешнеэкономическую деятельность территории. Для прогноза любого показателя, представленного в форме временного ряда, например объемов экспорта фармацевтической продукции из Италии, для отдельно взятого региона России, очевидно, необходимо либо обладать глубоким пониманием предметной области, либо учесть динамику изменчивости прочих показателей, характеризующих внешнеэкономическую деятельность регионов России, которые следует включить в модель как факторы и оценить, какие из них оказывают наибольшее влияние на результирующий показатель, т.е. решить задачу регрессии.

Поэтому основным вопросом является изучение адекватных правил формирования графа по объектам (группам товаров, странам, регионам) на основе большого количества признаков. При этом важно

избежать «проклятья размерности», т.е. избыточного описания специфики взаимодействия между элементами системы.

С решением этой задачи может справиться адекватный алгоритм понижения размерности, предназначенный для уменьшения сложности многомерной функции, описывающей взаимодействия элементов системы. Основной проблемой при реализации подобных подходов является необходимость обобщения данных, представленных в виде бинарных, категориальных и непрерывных числовых последовательностей, имеющих при этом различные статистические распределения. Помимо ряда существенных недостатков для большинства известных алгоритмов понижения размерности проекция множества компонентов набора данных, например для решения задачи регрессии, в общее абстрактное непараметрическое пространство, как и при использовании сверочных нейронных сетей, приводит к потере смысловой связи с источником происхождения этих данных и ухудшает интерпретируемость модели.

Наиболее перспективный среди подходов к описанию взаимодействия между элементами графа был предложен в 2018 г. Л. Макиннесом и успешно использован для нелинейного снижения размерности при построении моделей машинного обучения. Алгоритм Uniform Manifold Approximation and Projection (UMAP) [10; 11] позволяет выполнить равномерную аппроксимацию многообразия вариантов отображения точек (объектов) в многомерном пространстве признаков с поправкой на расстояние до ближайшего соседа каждой точки. На этой основе UMAP сформировано множество связей между объектами, т.е. простроен граф, описывающий структуру их взаимодействия.

Для алгоритма UMAP множество из ребер графа представляет собой нечеткое множество с функцией принадлежности, определяемой вероятностью существования ребра между двумя вершинами, что позволяет быстро создавать нечеткие наборы данных. UMAP использует экспоненциальное распределение вероятностей не только на евклидовых расстояниях (так, как это реализовано t-SNE [12]), а с возможностью использования любых метрик расстояний. Например, для категориальных данных может использоваться расстояние Гауэра [13], тогда как для порядковых данных используется Манхэттенское расстояние. Каждое представление и метрика могут использоваться для независимой генерации нечетких множеств, которые затем могут быть пересечены вместе, чтобы создать единый нечеткий набор.

Подробно работа алгоритма показана на рис. 1: для точек в многомерном пространстве, отображенных в системе координат двух главных компонент (c_1; c_2). Если одна из точек x_j при отображении в пространстве признаков оказывается достаточно далеко от остальных, то вычисляется расстояние до ее ближайшего соседа x_i $d(x_i, x_j)$, и это расстояние вычитается из расстояний до остальных точек, тем самым точка оказывается на расстоянии ближайшего соседа, что позволяет представить множество более компактно и выполнить кластеризацию точек. Однако при этом решается еще и задача нормировки распределений вероятностей для каждой точки. Для UMAP это условие нормировки связано с энтропией, где сумма вероятностей для всех точек N, составляющих множество, нормируется на энтропию по количеству ближайших соседей каждой точки.

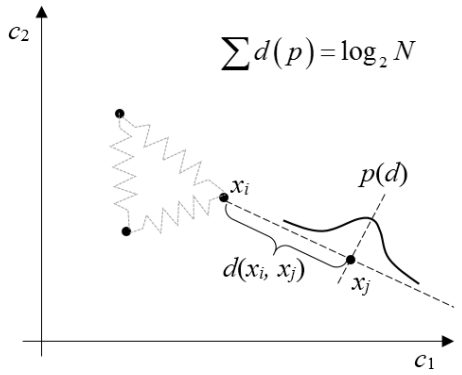


Рис. 1. Нормировка расстояний между точками UMAP *(авторские результаты)*

Оперируя вероятностями расстояний между точками в нечетком множестве, можно представить, что каждая из этих точек соединена с остальными воображаемыми «пружинами», в таком случае нормировка распределений вероятностей для расстояний между точками позволяет настраивать «жесткость» этих «пружин». Этот подход делает UMAP существенно продуктивнее алгоритма t-SNE при работе с массивами разнородных данных. Таким образом, распределение вероятностей для формирования весов связей имеет следующий вид:

$$p_{i|j} = e^{-\frac{d(x_i, x_j) - \rho_i}{\sigma_i}}, \quad (1)$$

где ρ является параметром, описывающим расстояние от каждой *i*-й точки данных до ее первого ближайшего соседа, что дает возможность варьировать метрическое расстояние для каждой точки данных. Отсутствие нормализации, а следовательно, и знаменателя степени в уравнении (1), сокращает время построения многомерного графа. UMAP определяет число ближайших соседей *k* следующим образом

$$k = 2^{\sum_i p_{ij}}. \quad (2)$$

Для достижения нормировки весов связей используют следующий подход:

$$p_{ij} = p_{i|j} + p_{j|i} - p_{i|j} \cdot p_{j|i}. \quad (3)$$

Нормировка или симметризация весов связей необходима на этапе группировки точек с локально изменяющимися метриками (с помощью параметра ρ), когда вес ребра графа между узлами *A* и *B* может быть не равен весу ребра между узлами *B* и *A*. При этом UMAP присваивает почти одинаковую координату в пространстве низкой размерности для всех точек, которые находятся близко друг к другу, что в итоге приводит к сверхплотно упакованным кластерам, часто наблюдаемым на графах уменьшения размерности UMAP. При этом UMAP присваивает начальные координаты в пространстве с низкой размерностью с использованием дискретного оператора Лапласа для конечного графа, в отличие от, например, случайной нормальной инициализации, используемой t-SNE.

За счет использования перекрестной энтропии в качестве функции потерь при оптимизации построения графа UMAP сохраняет как локальную, так и глобальную структуру связей между вершинами графа, в то же время значение перекрестной энтропии позволяет дать общее понимание, насколько далеко отстоят вершины друг от друга. Преимуществом UMAP является возможность построения графа по объектам на основе комплекса признаков, сформированных из различных показателей внешнеэкономической деятельности.

Однако для того чтобы сосредоточиться на анализе и выявлении групп вершин, связанных друг с другом, следует перейти к задаче выделения полных подграфов – клик [14; 15]. Кликой в неориентированном графе $G(V, E)$ называется подмножество вершин, каждая пара которых соединена ребром графа – это полный подграф первоначального графа, а максимальная клика – это клика, которая не

может быть расширена путем включения дополнительных смежных вершин. Выявление и анализ временных рядов показателей, описывающих внешнеэкономическую деятельность региона в виде количества вершин максимальной клики графа, т.е. кликового числа графа, позволяет выявить основные факторы, от которых зависит внешнеэкономическая деятельность региона.

Однако для реализации отбора признаков необходимо качественно и количественно оценить изменение структуры связей в графе при изменении масштаба графа путем включения и исключения новых ребер и вершин. Это утверждение позволяет выдвинуть гипотезу о том, что чем больше кликовое число графа N_q, тем больше степень локальной связанности графа, что в конечном счете влияет на глобальную структуру графа:

– при большом количестве вершин связей граф приобретает вид, близкий к симметричной структуре;

– при малом количестве связей и большом вершин для крупномасштабной сети возникает асимметричная доменная структура, несущая в себе основную информацию о структуре взаимосвязей между объектами и группами объектов (кластерами).

Для объективной оценки степени симметрии графа по количеству клик в нем следует соотнести это значение с размером графа. Эмпирически это соотношение выглядит следующим образом:

$$S_G = n / N_q , \tag{4}$$

где $n = N_G + M_G$ – размер сети, зависящий от количества узлов N_G и ребер M_G. Данный показатель для рассматриваемой задачи может быть определен как кликовый коэффициент симметрии структуры региональных внешнеэкономических связей.

Для построения оптимального графа на основе алгоритма UMAP необходимо варьировать количество ближайших соседей для каждой вершины и минимальное метрическое расстояние для установления связи между вершинами. Кликовый коэффициент симметрии графа позволяет оценить локальную связанность точек в построенном графе, формируя обобщенное представление о структуре сети с позиции наличия в ней кластеров.

Чем больше S_G, тем более полносвязным является граф, однако высокая степень локальной связанности графа не позволяет информативно описать структуру взаимосвязей между признаками. Поэтому

необходим показатель, который позволил бы оценить соответствие глобальной структуры графа реальным сетям.

На рис. 2 показан пример для расчета коэффициента Джини для оценки сбалансированности структуры графа. Дисбаланс описывается площадью ограниченной ломанной кривой Лоренца и рассчитывается по формуле:

$$GI = 1 - \sum_{k=1}^{n} \left(X_k - X_{k-1} \right) \cdot \left(Y_k + Y_{k-1} \right), \; I_G \in [0;1], \tag{5}$$

где n – количество точек; X_k – кумулятивная доля количества вершин графа; Y_k – кумулятивная доля количества связей графа. Для графа, у которого одна вершина имеет связи со всеми остальными, коэффициент Джини максимален и стремится к 1. В случае, когда все вершины графа имеют одинаковое количество связей, коэффициент Джини GI равен нулю. Чаще всего площадь занимает промежуточное значение, и данный критерий может быть использован при оптимизации структуры графа UMAP вместе с кликовым коэффициентом симметрии и позволяет дать оценку соответствия глобальной структуры графа реальным сетям.

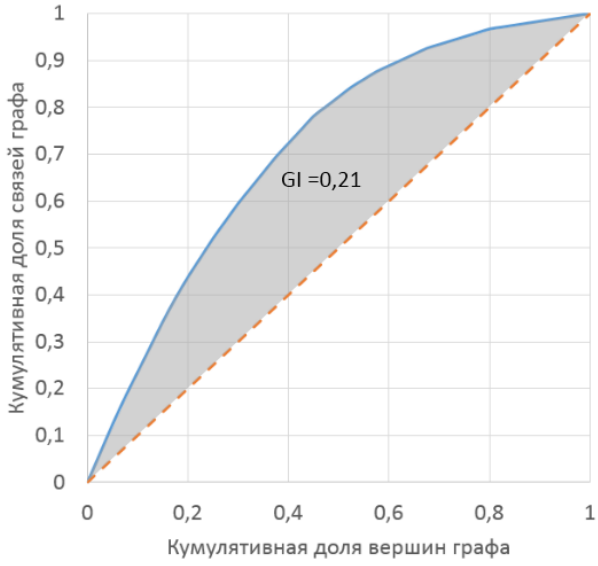


Рис. 2. Графическая интерпретация индекса Джини графа *(авторские результаты)*

Исследования показали [16], что сети, описывающие внешнеторговые отношения, также обладают плотными и очень неравномерными связями для большинства реальных крупномасштабных сетей, имеющих степенной закон распределения количества вершин от количества ребер в графа. Данное соотношение определяет коэффициент Джини и находится в пределах 0,15–0,2 в зависимости от структурных особенностей сети.

Задавшись пороговым значением коэффициента в указанных пределах Джини GI, задача оптимизации структуры графа решается путем максимизации $S_G \to \max$ в целях отображения связей межу вершинами, формирующими наиболее информативную структуру графа.

Информативность набора признаков, может быть оценена на основе показателя взаимной информации между парой узлов сети [17]

$$MI(x,y) = \log\left(\frac{p(x \mid y)}{p(x)}\right) = -\log p(x) - \left(-p(x \mid y)\right) = I(x) - I(x \mid y), \quad (6)$$

где $I(x)$ – информативность вершины *x*, характеризуемая количеством связей с ближайшими соседями, условная информативность вершины *y* при присоединении к вершине *x* с учетом общих соседей.

Таким образом, предлагаемый в работе подход может быть пошагово описан следующим образом:

1. Для сформированного набора данных формируется карта признаков на основе расстояний по косинусу [18], и визуализируется оптимизированный граф UMAP с учетом симметризации весов. Построенный граф является ориентированным и полносвязным, однако для отбора признаков и понимания их структуры необходимо оставить только те связи, которые вносят наибольший вклад в полноту описания характеристик предметной области, например, групп товаров, формирующих торговый профиль региона.

2. Выполняется бинарная классификация связей и отображение на графе только связей между объектами, не являющимися коллинеарными на основе косинусной меры сходства [18]. Косинусная мера сходства позволяет более адекватно отобразить точки в многомером пространстве признаков.

3. Для оптимизации структуры графа, т.е. включения / исключения связей между объектами, используется кликовый коэффициент симметрии и коэффициент Джини [19; 20], позволяющий сбалансировать глобальную структуру графа с позиции максимума информации для связанных признаков.

4. На основе полученной структуры графа необходимо вычислить максимальные клики для графа и далее выбрать перечень признаков, входящих в одну или несколько максимальных клик, имеющих максимум взаимной информации (*MI* – mutual information) [17; 21] для описания предметной области.

Моделирование

Модельные эксперименты с отображением графов различного вида в проекции пространства главных компонент для множества, состоящего из 200 точек, показаны на рис. 3 и вычислены на основе библиотеки umap для языка программирования python.

При большом количестве ближайших соседей для объединения точек (n_neighbors) и минимального расстояния между точками для объединения их в кластер (min_dist) отображение во всех случаях имеет симметричную по форме и хаотичную по расположению точек структуру.

Показатели n_neighbors и min_dist используются для контроля баланса между локальной и глобальной связностью точек в проекции пространства признаков. При увеличении n_neighbors UMAP соединяет все больше и больше соседних точек при построении графа в многомерном пространстве, что приводит к проекции, которая более точно отражает глобальную структуру данных. В свою очередь, min_dist имеет тенденцию к распределению точек в пространстве вложений, что приводит к уменьшению кластеризации данных и меньшему акценту на глобальную структуру.

Наличие в графе значительной доли «шумов» и случайных компонент, т.е. точек, имеющих неустойчивые слабые связи внутри множества, приводит к ухудшению информативности отображения точек в многомерном пространстве и не позволяет выявить глобальную структуру данных. Это говорит об отсутствии в системе устойчивых связей между признаками точек и невозможности оценить значимость межкластерных связей.

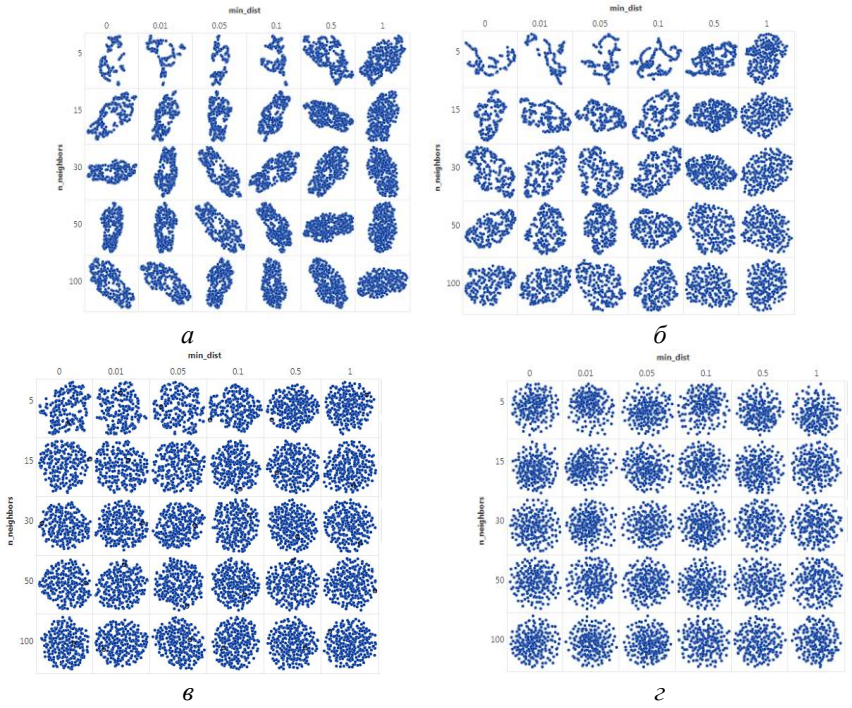


Рис. 3. Проекция наборов (*а – г*) из точек (*N* = 200) данных с различными значениями количества ближайших соседей для объединения точек (n_neighbors) и минимального расстояния между точками в кластер (min_dist) *(авторские результаты)*

Полученные результаты

На рис. 4 показана проекция ориентированного графа в пространстве двух главных компонент на основе косинусных расстояний – для упрощения отображения вершины не промаркированы (на примере структуры экспорта для Владимирской области за 2021 г.). В показанном случае граф n_neighbors =15, min_dist = 0,1 отображает глобальную структуру связей графа. Также UMAP реализует возможность отображения в пространстве нескольких главных компонент (рис. 5).

Показанные результаты формируют неориентированный граф, а для бинарной классификации для включения / исключения связей в граф необходимо выполнить оптимизацию порогового значения весов связей между вершинами графа и оставить только наиболее значимые связи с точки зрения решаемой задачи. При этом критерием оптимизации является максимум кликового коэффициента симметрии

при значении коэффициента Джини графа не менее заданного ($I_G \geq 0,2$).

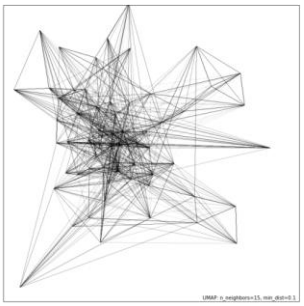
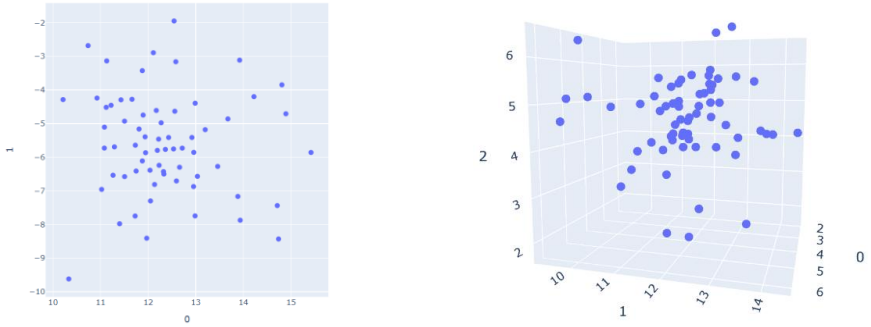


Рис. 4. Проекция графа в пространстве двух главных компонент на основе косинусных расстояний *(авторские результаты)*



а *б*

Рис. 5. Проекции продуктов в пространстве двух (*а*) и трех (*б*) главных компонент, построенных на основе UMAP для экспорта Владимирской области за 2021 г., с оптимизацией относительной локальной плотности данных на основе косинусных расстояний *(авторские результаты)*

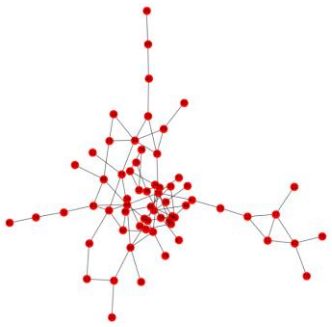
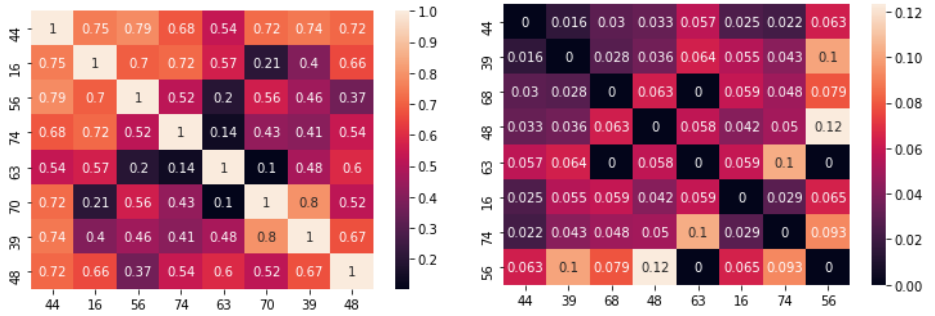


Рис. 6. Проекция продуктов в пространстве двух главных компонент, построенных на основе UMAP для экспорта Владимирской области за 2021 г., с оптимизацией относительной локальной плотности данных на основе косинусных расстояний *(авторские результаты)*



а *б*

Рис. 7. Тепловые карты парных коэффициентов корреляции (*а*) и косинусных мер сходства (*б*) отобранных признаков *(авторские результаты)*

Сформированный граф имеет максимальное значение $S_G = 2,76$ при значении индекса Джини графа $S_G = 0,22$. Первые пять максимальных клик полученного графа имеют вид: ([44, 48, 63], [44, 70, 39], [44, 39, 68], [44, 16, 56], [44, 16, 74]). Это означает, что товарные группы в соответствии с номенклатурой ТНВЭД с номерами [44, 48, 63, 70, 39, 68, 16, 56, 74] составляют основной набор предикторов для построения модели. Полученные признаки не имеют сильной парной корреляции по шкале Чеддока (рис. 7), но между некоторыми из них прослеживается высокая корреляционная зависимость, и данный набор показателей имеет максимум взаимной информации *MI* при минимуме признаков, что максимально уменьшает искажения при описании структуры региональных внешнеэкономических связей; они используются для описания факторов изменчивости при построении прогностических моделей показателей внешнеэкономической деятельности региона.

Следует отметить что дальнейшее увеличение взаимной информации при дополнении перечня признаков вершинами из последующих клик в модельном примере не приводило к значительному росту показателя *MI*. При этом мера косинусного сходства для отобранный признаков невелика, что также означает, что векторы не являются коллинеарными в метрическом пространстве.

Заключение

После выбора оптимальной клики на основе взаимной информации был выявлен перечень продуктов, состоящий из 9

вершин, обладающих отсутствием мультиколлинеарности, и также имеющие максимальный показатель взаимной информации при оптимальной структуре графа, обеспечивающей конструирование множества признаков, пригодных для построения прогностической модели как классификации объектов, так и регрессии.

Применение описанного в работе подхода позволило устранить мультиколлинеарность признаков, осуществить отбор показателей, расширить возможности использования имеющегося набора данных за счет включения новых показателей, вносящих в модель новую полезную информацию о предметной области.

Использование отобранных признаков на основе анализа в проекции главных компонент позволит преодолеть «проклятье» размерности, устранить «шумы» и снизить переобучение модели.

Рассмотренный в работе метод отбора признаков рационально использовать для построения интерпретируемых прогностических моделей показателей ВЭД и как один из способов понижения размерности пространства признаков модели.

Список литературы

1. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Appli-cations in R. – Publisher: Springer, 2013. – 436 p.

2. Шитиков В.К., Мастицкий С.Э. Классификация, регрессия, алгоритмы Data Mining с использованием R. 2017 [Электронный ресурс]. – URL: https://github.com/ranalytics/data-mining (дата обращения: 24.02.2021)

3. The atlas of economic complexity: Mapping paths to prosperity / R. Hausmann, C.A. Hidalgo, S. Bustos [et al.] // Mit Press, 2014.

4. Filimonova M., Kislyakov A., Tikhonyuk N. Structural and Dynamic Modelling of the Regions' Foreign Trade Profile Based on Graph Cluster Analysis // STRATEGICA: Shaping the Future of Business and Economy. – Bucharest, 2021. P. 34–49.

5. Шитиков В.К. Моделирование корреляционных связей в сообществе с помощью сетей [Электронный ресурс]. – URL: http://www.ievbras.ru/-ecostat/Kiril/R/Blog/14_QGraph.pdf (дата обращения: 08.09.2020).

6. A study of effects of multicollinearity in the multivariable analysis / W. Yoo, R. Mayberry, S. Bae, K. Singh, QP He, Jr JW. Lillard // International journal of applied science and technology. – 2014. – № 4(5). – P. 9–19.

7. Kislyakov A., Tikhonuyk N. Principles for Development of Predictive Stability Models of Social and Economic Systems on the basis of DTW // E3S Web

of Conferences. – 2020. – Vol. 208. – P. 08001. DOI: 10.1051/e3sconf/-202020808001

8. Федеральная таможенная служба РФ [Электронный ресурс]. Таможенная статистика внешней торговли РФ, 2021. офиц. сайт. – URL: http://stat.customs.ru (дата обращения: 21.06.2021).

9. Об утверждении методологии ведения статистики взаимной торговли товарами государств – членов Евразийского экономического союза и методологии ведения таможенной статистики внешней торговли товарами государств – членов евразийского экономического союза // Коллегия евразийской экономической комиссии. Решение от 25 декабря 2018 г. – 2018. – № 210.

10. McInnes L., Healy J., Melville J. Umap: Uniform manifold approximation and projection for dimension reduction // arXiv preprint arXiv. – 2018. – Vol. 1802. – P. 03426.

11. Dimensionality reduction for visualizing single-cell data using UMAP / Becht E. [et al.] // Nature biotechnology. – 2019. – Vol. 37, №. 1. – P. 38–44.

12. Van der Maaten L., Hinton G. Visualizing data using t-SNE //Journal of machine learning research. – 2008. – Vol. 9, №. 11.

13. Franklin J. The elements of statistical learning: data mining, inference // and prediction. The Mathematical Intelligencer. – 2003. – № 27. – P. 83–85. DOI: 10.1007/BF02985802

14. Henniab K., Mezghani N., Gouin-Vallerand C. Unsupervised Graph-Based Feature Selection Via Subspace and PageRank Centrality, [Электронный ресурс]. – URL: https://bit.ly/2HGON5B (дата обращения: 21.01.2022).

15. Dicks W., Dunwoody M. J. Groups Acting on Graphs, Cambridge Studies in Advanced Mathematics. – 1989. – Vol. 17.

16. Strogatz S.H. Syncing: how order arises from chaos in the universe, nature and everyday life. – Hachette Books, 2012. – 352 p.

17. Kumar P., Sharma, D. A potential energy and mutual information based link prediction approach for bipartite networks // Scientific Reports. – 2020. – № 10. – P. 20659.

18. Форман Дж. Много цифр: Анализ больших данных при помощи Excel. – М.: Альпина Паблишер, 2016. – 464 с.

19. Kislyakov A.N. Structuring advertising campaign costs considering the asymmetry of users' interests // Business Informatics. – 2020. – Vol. 14, No 4. – P. 7–18. DOI: 10.17323/2587-814X.2020.4.7.18

20. Biró, T. S., Néda, Z. Gintropy: Gini Index Based Generalization of Entropy // Entropy. – 2020. – № 22(8). – P. 879. DOI:10.3390/e22080879

21. Tan F., Xia Y., Zhu B. Link Prediction in Complex Networks: A Mutual Information Perspective // PLOS ONE. – 2014. – Vol. 9. – P. e107056.

References

1. James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Appli-cations in R. Publisher, Springer, 2013, 436 p.

2. Shitikov V. K., Mastitsky S. E. Klassifikaciya, regressiya, algoritmy Data Mining s ispol'zovaniem R [Classification, regression, Data Mining algorithms using R]. 2017, E-book, access address. Available at: https://github.com/-ranalytics/data-mining

3. Hausmann R., Hidalgo C. A., Bustos S. et al. The atlas of economic complexity: Mapping paths to prosperity. Mit Press, 2014.

4. Filimonova M., Kislyakov A., Tikhonyuk N. Structural and Dynamic Modelling of the Regions' Foreign Trade Profile Based on Graph Cluster Analysis. STRATEGICA: Shaping the Future of Business and Economy. Bucharest, October 2021, p. 34-49.

5. Shitikov V.K. Modeling of correlations in the community using networks. Available at: http://www.ievbras.ru/ecostat/Kiril/R/Blog/14_QGraph.pdf (accessed: 08 September 2020)

6. Yoo W., Mayberry R., Bae S., Singh K., He Q.P., Lillard Jr J.W. A study of effects of multicollinearity in the multivariable analysis. *International journal of applied science and technology,* 2014, Oct 4(5), pp. 9-19.

7. Kislyakov A., Tikhonuyk N. Principles for Development of Predictive Stability Models of Social and Economic Systems on the basis of DTW. E3S Web of Conferences, 2020, Vol. 208, pp. 08001. DOI: 10.1051/e3sconf/202020808001

8. Federal Customs Service of the Russian Federation: ofic. website – Customs statistics of foreign trade of the Russian Federation, 2022. Available at: http://stat.customs.ru.

9. On approval of the methodology for maintaining statistics of mutual trade in goods of the member States of the Eurasian Economic Union and the methodology for maintaining customs statistics of foreign trade in goods of the member States of the Eurasian Economic Union. *Board of the Eurasian Economic Commission.* Decision of December 25, 2018, no 210.

10. McInnes L., Healy J., Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv;* 2018. no 1802. pp. 03426.

11. Becht E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology*, 2019, Vol. 37, no 1, pp. 38-44.

12. Van der Maaten L., Hinton G. Visualizing data using t-SNE. *Journal of machine learning research*, 2008, Vol. 9, no 11.

13. Franklin J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer,* 2003, no. 27, pp. 83–85. DOI: 10.1007/BF02985802

14. Henniab K., Mezghani N., Gouin-Vallerand C. Unsupervised Graph-Based Feature Selection Via Subspace and PageRank Centrality. Available at: https://bit.ly/2HGON5B (accessed 21 Januari 2022).

15. Dicks W., Dunwoody M.J. Groups Acting on Graphs. *Cambridge Studies in Advanced Mathematics*, 1989. vol. 17,

16. Strogatz S.H. Syncing: how order arises from chaos in the universe, nature and everyday life. Hachette Books, 2012, 352 p.

17. Kumar P., Sharma, D. A potential energy and mutual information based link prediction approach for bipartite networks. *Scientific Reports*, 2020, no 10, pp. 20659.

18. Foreman J. A lot of numbers: Big data analysis using Excel. Moscow, Alpina Publisher, 2016, 464 p.

19. Kislyakov A.N. Structuring advertising campaign costs considering the asymmetry of users' interests. *Business Informatics*, 2020, vol. 14, no 4, pp. 7–18. DOI: 10.17323/2587-814X.2020.4.7.18

20. Biró, T. S., Néda, Z. Gintropy: Gini Index Based Generalization of Entropy. *Entropy*, 2020, vol. 22(8), pp. 879. DOI:10.3390/e22080879

21. Tan F., Xia Y., Zhu, B. Link Prediction in Complex Networks: A Mutual Information Perspective. *PLOS ONE,* 2014, no. 9, pp. e107056.

Статья получена: 24.01.2022

Статья одобрена: 28.02.2022

Принята к публикации: 18.03.2022

Финансирование. *Исследование не имело спонсорской поддержки.*

Конфликт интересов. *Авторы заявляют об отсутствии конфликта интересов.*

Сведения об авторе

Кисляков Алексей Николаевич (Владимир, Россия) – кандидат технических наук, доцент, доцент кафедры информационных технологий, Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации, Владимирский филиал (600017, г. Владимир, ул. Горького, д. 59а, e-mail: ankislyakov@mail.ru).

About the author

Alexey N. Kislyakov (Vladimir, Russian Federation) – Ph. D. in of Engineering, Associate Professor, Associate Professor of the Department of Information

Technology, Russian Academy of National Economy and Public Administration under the President of the Russian Federation, Vladimir branch (59a, Gorky str., Vladimir 600017, e-mail: ankislyakov@mail.ru).

Библиографическое описание статьи согласно ГОСТ Р 7.0.100–2018:

Кисляков, А. Н. Отбор признаков для использования в моделях предиктивной аналитики внешнеэкономической деятельности регионов / А. Н. Кисляков. – текст : непосредственный. – DOI: 10.15593/2499-9873/2022.1.09 // Прикладная математика и вопросы управления = Applied Mathematics and Control Sciences. – 2022. – № 1. – С. 176–195.

Цитирование статьи в изданиях РИНЦ:

Кисляков, А. Н. Отбор признаков для использования в моделях предиктивной аналитики внешнеэкономической деятельности регионов / А. Н. Кисляков // Прикладная математика и вопросы управления. – 2022. – № 1. – С. 176–195. DOI: 10.15593/2499-9873/2022.1.09

Цитирование статьи в references и международных изданиях

Cite this article as:

Kislyakov A.N. Features selection for use in predictive analytics models of regions foreign economic activity. *Applied Mathematics and Control Sciences*, 2022, no. 1, pp. 176–195. DOI: 10.15593/2499-9873/2021.2.09 (*in Russian*)