Алгоритмы машинного обучения на основе статистических данных о расходах и доходах населения с учетом их дифференциации

Е.А. Макарова
Факультет информатики робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
e-mail: ea-makarova@mail.ru

Е.Ш. Закиева
Факультет информатики робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
zakievae@mail.ru

Н.В. Хасанова
Факультет информатики робототехники
Уфимский государственный авиационный технический университет
Уфа, Россия
e-mail: khasanova.nv@mail.ru

Аннотация[1]

Рассматриваются вопросы разработки алгоритмов машинного обучения на основе статистических данных о расходах и доходах населения с учетом их дифференциации по регионам Российской Федерации. Проведен интеллектуальный анализ данных о дифференциации доходов населения регионов Российской Федерации с учетом их расходов. Сформированы закономерности кластеризации регионов на основе данных о дифференциации доходов населения.

Ключевые слова: алгоритм; уровень дифференциации доходов; расходы; машинное обучение.

Введение

Современная социально-экономическая ситуация в России характеризуется недостаточно высокими темпами развития и выпуска валового внутреннего продукта (ВВП), неудовлетворительным уровнем и качеством жизни населения, растущими миграционными потоками, достаточно высокой межрегиональной дифференциацией.

Расширение самостоятельности регионов, развитие бюджетного федерализма усиливают значимость региональной экономической политики, и направлено на повышение уровня и качества жизни населения, и снижение дифференциации доходов населения. Оценка уровня жизни является аналитической основой для принятия решений по выделению субсидий субъектам РФ, для управления социальными процессами и внесения корректив в политику в социальной сфере, пересмотра размеров затрат по статьям расходов.

Вопросы сравнения показателей уровня жизни населения различных субъектов РФ, оценки степени их дифференциации, определения направления динамики различных интегральных свойств уровня жизни в каждом из субъектов, причин наблюдаемых сдвигов и корректировки ключевых направлений в совершенствовании социально-экономической политики требуют разработки специальной методологии межрегионального анализа уровня жизни населения субъектов РФ и их динамики.

Процедура анализа данных о доходах и расходах населения с использованием алгоритмов машинного обучения

Исследование уровня жизни населения или его отдельных групп основывается на системе различных показателей, которые традиционно подразделяются на общие и частные. Общие показатели характеризуют общие достижения экономического развития общества (национальный доход, фонд потребления, величина жилищного фонда и т.д.). Частные показатели связаны с отдельными территориями, отдельными группами населения, они более детализированы.

Разработка процедуры анализа статистических данных о расходах и доходах населения с учетом

Труды Восьмой всероссийской научной конференции "Информационные технологии интеллектуальной поддержки принятия решений", 6-9 октября, Уфа-Ставрополь, Ханты-Мансийск, Россия, 2020

дифференциации доходов населения ведется на основе алгоритмов машинного обучения с использованием взаимосвязанного применения методов компонентного и кластерного анализа [1-4]. Процедура имеет следующие особенности:

- обеспечивает структурирование многочисленных факторов, определяющих уровень жизни населения;
- учитывает не только количественные, но и качественные факторы;
- основывается на анализе статистических данных, при этом обеспечивая объективность, которая означает, что обнаруженные закономерности будут полностью соответствовать действительности;
- обеспечивает выявление закономерностей кластеризации регионов с учетом дифференциации доходов населения.

Процедура анализа статистических данных о расходах и доходах населения с учетом их дифференциации состоит в том, что они предполагают многократное последовательное применение алгоритмов машинного обучения при выборе различных параметров методов для нескольких обучающих выборок, полученных путем структуризации статей доходов и расходов населения с учетом дифференциации по доходам, что позволяет выделить кластеры и визуализировать их в различных проекциях, а также сформировать их отличительные характеристики [5-8].

Разработан алгоритм формирования итогового рейтинга на основе применения компонентного анализа данных с использованием матрицы значений интегральных признаков и матрицы весовых коэффициентов. Особенности алгоритма формирования рейтинга регионов по уровню жизни состоит, во-первых, в том, что он одновременно учитывает результаты компонентного анализа данных всех анализируемых признаков, объединенных в три группы; и, во-вторых, предполагает построение новой интегральной выборки на основе результатов компонентного анализа трех групп признаков и проведение повторного компонентного анализа для нее.

В результате проведения компонентного анализа выборки с использованием автоматизированной информационной системы, включающей разработанные алгоритмы анализа данных об уровне жизни населения регионов, выявлены кластеры регионов РФ, различающихся по уровню дифференциации доходов населения с учетом их расходов.

Первым типом анализируемой выборки является выборка, характеризующая состояние доходов населения регионов РФ. Цель анализа сформированной выборки заключается в выявлении кластеров регионов РФ, различающихся по уровню дифференциации доходов населения (по основным видам доходов) с учетом занятости.

В результате проведения компонентного анализа подготовленной выборки получена сводная характеристика трех компонент с указанием собственных значений. Полученные результаты говорят о том, что первые три главные компоненты описывают 89,06 % дисперсии исходных данных, что достаточно для анализа пространственного распределения объектов. График для собственных значений компонент представлен на рисунке 1.

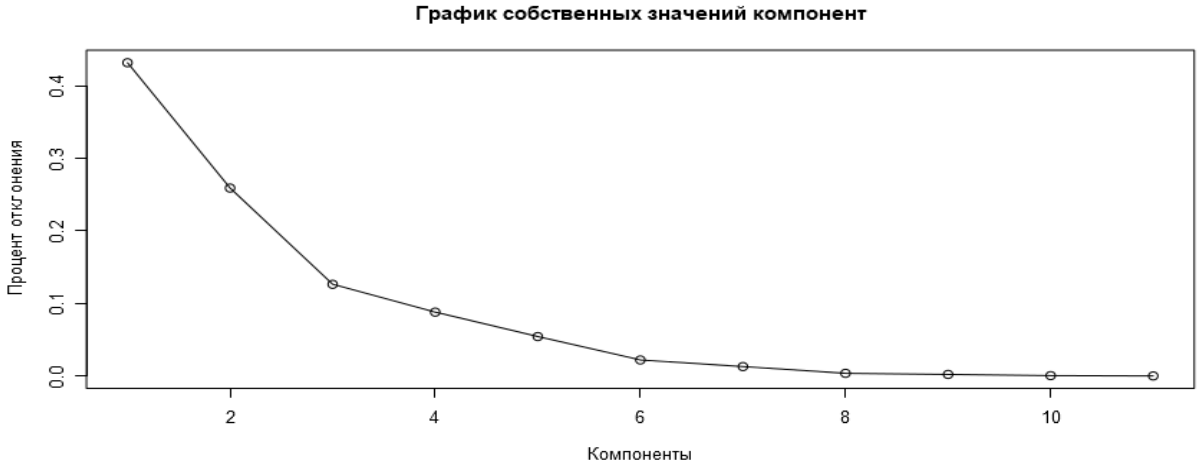

Рис. 1. График собственных значений компонент для выборки «Доходы»

По результатам выполненного компонентного анализа построены три главные компоненты. Первая компонента характеризует уровень трудовых доходов населения. В области больших значений первой компоненты находятся следующие регионы-лидеры по уровню трудовых доходов: г. Москва, Магаданская область, Чукотский автономный округ, Камчатский край, Тюменская область, Сахалинская область, г. Санкт-Петербург (рисунок 2). В области малых значений первой компоненты расположены

отстающие по уровню трудовых доходов регионы: Республика Дагестан, Кабардино-Балкарская Республика, Карачаево-Черкесская Республика, Республика Ингушетия.

Вторая компонента характеризует – уровень социальных выплат с учетом скрытых доходов населения. В области больших значений второй компоненты находятся следующие регионы-лидеры по уровню нетрудовых доходов: г. Москва, Московская область, республика Дагестан, Краснодарский край, Республика Башкортостан, Республика Татарстан, г.Санкт-Петербург и другие. В области малых значений второй компоненты расположены регионы с низким уровнем социальных выплат с учетом скрытых доходов населения: Республика Калмыкия, Республика Тыва, Республика Карелия, Камчатский край.

Наблюдается следующая тенденция: с увеличением среднегодовой численности занятых уменьшаются социальные выплаты, и увеличиваются другие доходы, в том числе «скрытые».

Третья компонента характеризует уровень доходов от предпринимательской деятельности с учетом доходов от собственности согласно следующей закономерности: с увеличением доходов от собственности, уменьшаются доходы от предпринимательской деятельности, и наоборот. Интересен тот факт, что в г. Москва доходы от собственности преобладают над доходами от предпринимательской деятельности.

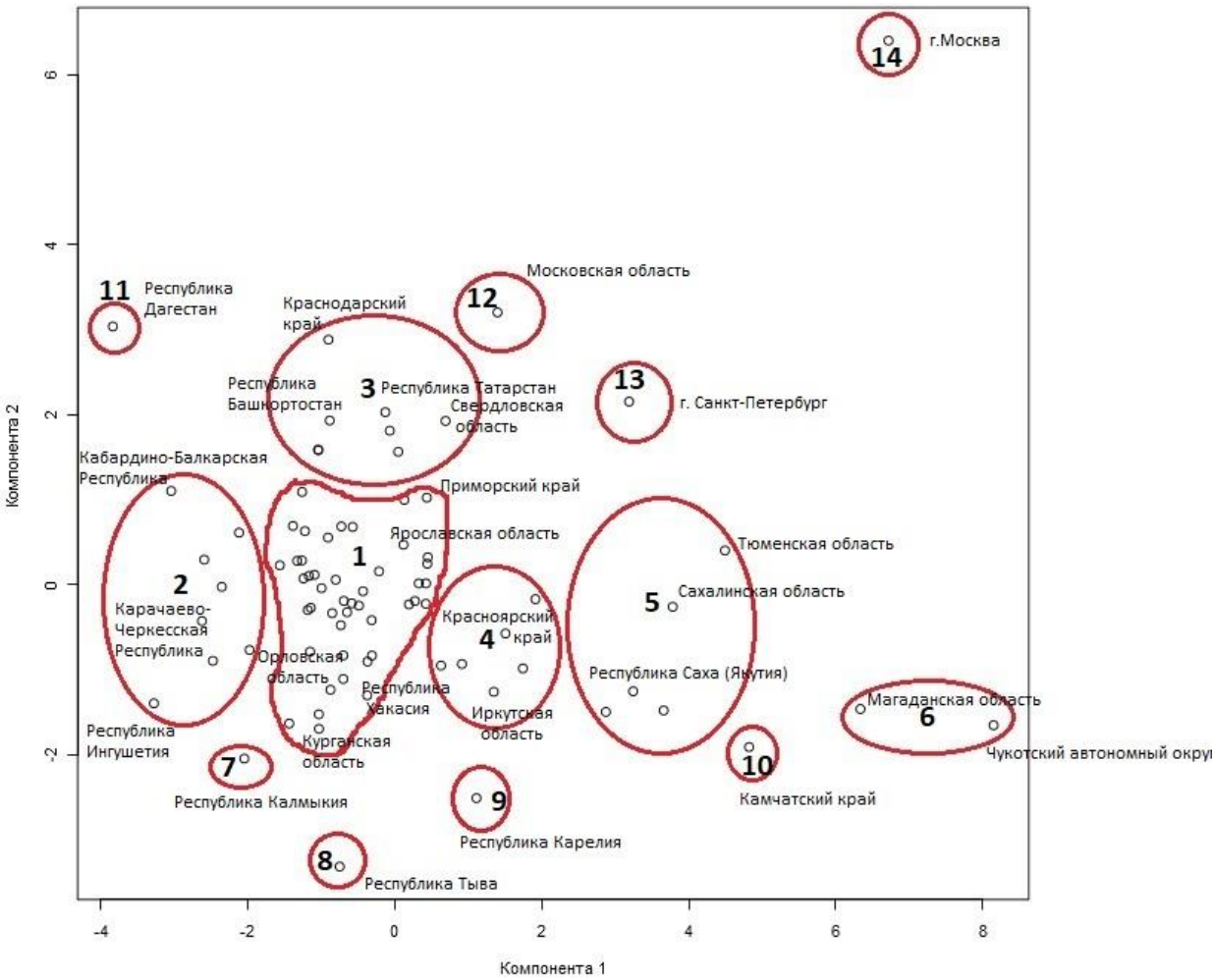

Рис. 2. Проекция регионов РФ в пространстве первой и второй компоненты

Вторым типом анализируемой выборки является выборка, характеризующая уровень расходов населения регионов РФ. Цель её анализа заключается в выявлении кластеров регионов РФ, различающихся по уровню расходов населения по основным статьям затрат: расходы на оплату услуг; расходы на покупку непродовольственных товаров; расходы на покупку продуктов питания; потребительские расходы в целом; индексы потребительских цен; расходы на одежду и обувь; расходы на жилищно-коммунальные услуги; расходы на транспорт; расходы на образование.

Первая компонента характеризует уровень расходов по их основным видам (продовольственные, непродовольственные, услуги), вторая – уровень расходов на образование с учетом инфляции, третья компонента – уровень расходов на транспорт с учетом инфляции.

В ходе выполнения компонентного анализа данных для выборки «Расходы» с помощью разработанной автоматизированной информационной системы анализа данных об уровне жизни населения регионов выделено всего семь кластеров, из них два кластера содержат по два региона, четыре кластера являются малочисленными, и один кластер – многочисленный. Самый большой уровень расходов наблюдается в регионах г. Москва, Московская область, Хабаровский край, Сахалинская область, Камчатский край, Магаданская область, Чукотский автономный; самый низкий уровень расходов наблюдается в регионах Республика Ингушетия, Республика Тыва, Брянская область, Республика Дагестан, Республика Калмыкия, Алтайский край.

Третьим типом анализируемой выборки была выборка, характеризующая уровень дифференциации доходов населения регионов РФ с учетом их расходов на продукты питания. Выполненный кластерный анализ показал, что в регионах, у которых двадцатипроцентная группа населения имеет большую долю доходов, уровень дифференциации выше. Анализ расположения регионов в пространстве в плоскости первой и второй компонент позволил заключить, что большинство регионов занимают центральное положение. Это говорит о средней степени дифференциации населения по доходам (как по двадцатипроцентным, так и десятипроцентным группам населения). Среднее расположение по второй компоненте говорит о сбалансированности расходов на питание. Обращает внимание также следующий интересный момент: выделяются два региона по первой компоненте, которые занимают крайнее правое положение, что свидетельствует о низком уровне дифференциации (г. Севастополь и Республика Крым).

Для расчета результирующего рейтинга использовалась интегральная выборка, полученная на основании матрицы весовых коэффициентов при проведении компонентного анализа предыдущих выборок о доходах, расходах и дифференциации доходов населения регионов РФ. Разработанный алгоритм построения результирующего рейтинга на основе статистических данных о расходах и доходах населения регионов с учетом их дифференциации позволил получить необходимые значения интегрального признака и построить результирующий рейтинг регионов по уровню жизни с учетом дифференциации доходов населения Регионами-лидерами по уровню жизни согласно рейтингу стали: г. Москва, г. Санкт-Петербург, Тюменская область, Республика Татарстан, Краснодарский край, Московская область, Калининградская область.

Заключение

С помощью разработанной процедуры анализа статистических данных о расходах и доходах населения с учетом дифференциации доходов населения на основе алгоритмов машинного обучения выполнен компонентный и кластерный анализ данных по выборке, характеризующей уровень доходов населения регионов РФ. Построенные кластеры регионов различаются по уровню доходов населения в их основных видах – трудовых, социальных выплат и доходов от предпринимательской деятельности.

Выделены различные по составу кластеры – малочисленные и многочисленные, в каждом из которых преобладает определённый вид доходов – трудовые, социальные выплаты с учетом скрытых доходов населения или доходы от предпринимательской деятельности. Выявлены также закономерности, показывающие, например, что при уменьшении социальных выплат увеличивается доля «скрытых» доходов.

Выполнен компонентный и кластерный анализ данных по выборке, характеризующей уровень расходов населения регионов РФ. Построенные кластеры регионов различаются по уровню расходов населения по их основным видам (расходы на продовольственные товары, непродовольственные товары, услуги) с учетом инфляции. Выделены различные по составу кластеры – малочисленные и многочисленные, в каждом из которых преобладает определённый вид расходов.

Выполнен компонентный и кластерный анализ данных по выборке, характеризующей уровень дифференциации доходов населения регионов РФ с учетом их расходов. Построенные кластеры регионов различаются по уровню дифференциации доходов. Показано, что наибольший уровень дифференциации имеет место в регионах с высоким уровнем доходов и наоборот.

Выполнен расчет итогового рейтинга регионов РФ по уровню жизни населения. Особенность алгоритма состоит в вычислении итогового рейтинга регионов путем проведения компонентного анализа интегральной выборки. Анализ значимости главных компонент позволил остановиться на выборе только одной – первой главной компоненты. С помощью рассчитанного показателя уровня жизни все регионы упорядочены в одномерном пространстве в порядке убывания, сформирован рейтинг регионов. Регионами-лидерами по уровню жизни согласно рейтингу стали: г. Москва, г. Санкт-Петербург, Тюменская область, Республика Татарстан, Краснодарский край, Московская область, Калининградская область.

Благодарности

Работа выполнена при поддержке гранта РФФИ № 20-08-00796 «Интеллектуальное управление промышленным комплексом как динамическим многоагентным объектом на основе методов когнитивного моделирования и машинного обучения».

Список используемых источников

1. Гузаиров М. Б., Дегтярева И. В., Макарова Е. А. Расходы населения регионов российской федерации на покупку продуктов питания: компонентный и кластерный анализ // Экономика региона, – 2015. – № 4 (44). – С.145-158.
2. Ильясов Б. Г., Макарова Е. А., Закиева Е. Ш., Гиздатуллина Э. С. Анализ данных о доходах населения регионов РФ с учетом социальных трансфертов: метод главных компонент // Фундаментальные исследования. Издательство: Издательский Дом "Академия Естествознания", – 2018. – № 4. – С. 69-74.
3. Ильясов Б. Г., Макарова Е. А., Закиева Е. Ш., Габдуллина Э. Р., Гиздатуллина Э. С. Компонентный и нейросетевой анализ доходов населения с учетом его численности // Математические модели современных экономических процессов, методы анализа и синтеза экономических механизмов. Актуальные проблемы и перспективы менеджмента организаций в России: сб. ст. XI Всерос. науч.-практ. конф. Вып. 11. – Самара: Изд-во СамНЦ РАН, – 2017. – С.27-33.
4. Ильясов Б. Г., Макарова Е. А., Закиева Е. Ш., Габдуллина Э. Р. Метод формирования рейтинга регионов в сфере образования, труда, инноваций на основе интеллектуальных алгоритмов. // Информационные технологии интеллектуальной поддержки принятия решений: V Всеросс. конф., 16-19 мая, Уфа, Россия, 2017.
5. Макарова Е.А., Ровнейко Н.И., Хасанова Н.В. Интеллектуальный анализ отраслевой структуры реального сектора экономики на основе данных межотраслевого баланса // Управление экономикой: методы, модели, технологии: материалы XVIII Международной научной конференции. – Уфа: РИК УГАТУ, – 2018. – С.358-361.
6. Ильясов Б.Г., Хасанова Н.В. Гиздатуллина Э.С. Интеллектуальные алгоритмы анализа данных об инвестиционных процессах регионального уровня // Управление экономикой: методы, модели, технологии: материалы XVI Международной научной конференции. – Уфа: РИК УГАТУ, – 2016. – С. 463-470.
7. Макарова Е.А., Герасимова И.Б., Закиева Е.Ш., Масленникова Ю.А. Компонентный и кластерный анализ отраслевой структуры реального сектора экономики // Информационные технологии интеллектуальной поддержки принятия решений (ITIDS'2018): Труды VI Всероссийской конференции (с приглашением зарубежных ученых). – Уфа: РИК УГАТУ, – 2018. Том 2. С.184-189.
8. Макарова Е.А., Хасанова Н.В. Габдуллина Э.Р. Анализ инвестиционной активности регионов РФ: деревья решений и нейронные сети Кохонена. // Системный анализ в проектировании и управлении: XXI Международная научно-практическая конференция. – СПб.: Изд-во Политехн. ун-та, – 2017. – Ч.2. –С. 91-98.