Введение.

Классификация субъектов Российской Федерации является важным инструментом для упорядочения и систематизации административно-территориального деления страны.

Исследуя результаты, полученные экспертами, появляется возможность определить типовые регионы, разработать целенаправленные стратегии развития для каждой группы, что способствует более эффективному управлению региональным развитием. [18]

Более того, сформированные группы субъектов могут свидетельствовать о наличии в стране пространственной поляризации (неравномерность социально-экономического развития территории, выражающаяся в разных уровнях социально-экономического развития ее регионов), которая может возникать вследствие перераспределения экономических ресурсов, доходов и расходов. [20]

В настоящее время инновационное развитие страны и отдельных регионов является одной из наиболее приоритетных задач государства. Во многих работах экономистов для типологизации регионов используется именно предрасположенность к научно-техническому прогрессу. [24] Инновационные характеристики также важны для оценки стратегической конкурентоспособности регионов. [23]

Реализовать подобные разбиения можно, используя методы машинного обучения, а именно - кластеризацию. Процедура подходит для анализа данных, позволяет группировать объекты на основе их сходства. Применение данной методики к субъектам Российской Федерации даёт возможность выделить группы регионов, имеющие схожие характеристики в рассматриваемой области.

Кластерный анализ используется в выявлении проблемных регионов и исследовании факторов, оказывающих на них негативное влияние. [19] Это, в свою очередь, будет сигнализировать о необходимости принятия мер по стимулированию развития отстающих регионов. Исследование вопроса кластеризации субъектов РФ по показателям, относящимся к инновационной деятельности описано в статьях [21, 22]. Отталкиваясь от рассмотренных статей, можно предположить, что существуют статичные группы субъектов по своим экономическим и инновационным показателям.

Перед реализацией разбиения необходимо качественное конструирование признакового пространства, максимально точно описывающего свойства рассматриваемых субъектов в выбранной области исследования. (уровень жизни, экономика, социально-экономический аспект, миграция и т д) Важным аспектом является выбор алгоритма кластеризации. Он осуществляется на основе характеристик данных и на результате, которого стремится достичь исследователь.

Чёткого и правильного разбиения не существует. Есть возможность оценить результат математически, используя определённые метрики, основанные на внутрикластерной схожести и межкластерном различии. Либо интерпретировать качество полученных групп основываясь на чьем-либо экспертном мнении, контекстной информации. Для того, чтобы приблизиться к классификации, полученной экспертами, необходимо исследовать наиболее популярные алгоритмы кластеризации, реализовать машинный эксперимент, осуществив подбор параметров. Также необходимо понять, какое количество информации использовать.

Возникает вопрос - возможно ли, пользуясь только имеющимися данными, методами и алгоритмами достичь результата, который получен и обоснован экономистами?

Постановка задачи

Были рассмотрена статья с результатами кластеризации регионов России по экономическим и инновационным показателям. [25] Также найдены сформированные экспертами группы регионов по общему уровню развития инноваций. [26]

Так, в первой главе необходимо сформировать набор данных по субъектам, рассмотреть имеющиеся программы кластеризации, метрики для оценки результатов и провести машинный эксперимент с имеющимися данными. Прийти к выводам, какой из методов позволяет наиболее точно приблизиться к нужному результату.

При рассмотрении иерархических методов кластеризации возникает проблема отсутствия чёткого критерия для определения оптимального числа кластеров. А так как подход к решению данной задачи основывается именно на кластеризации, а не на классификации (то есть мы предполагаем, что заранее не знаем нужное число групп, на которые будут разбиты субъекты) - необходим некий метод, основанный на анализе данных об объектах. Реализация подобного критерия описана в статье Орехова А.В.[17]. Для него потребуется программная реализация.

Во время проведения машинных экспериментов возникает желание рассмотреть изменение характеристик регионов в динамике и провести кластеризацию временных рядов. Также исследовав полученные кластерные группы и проведя более детальные исследования имеющихся данных, возможно рассмотрение наиболее подходящего метода кластеризации, основанного на вычислении внутрикластерной плотности.