МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н.Г. ЧЕРНЫШЕВСКОГО»**

Кафедра социальной информатики

ИЕРАРХИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ЭЛЕКТОРАЛЬНЫХ ДАННЫХ

(автореферат бакалаврской работы)

студента 4 курса 451 группы
направления 09.03.03 - Прикладная информатика
профиль Прикладная информатика в социологии
Социологического факультета
Зайцева Валерия Александровича

Научный руководитель
кандидат физико-математических наук, доцент _____ Л.Б. Тяпаев
подпись, дата

Зав. кафедрой
кандидат социологических наук, доцент _____ И.Г. Малинский
подпись, дата

Саратов 2022

ВВЕДЕНИЕ

Актуальность проблемы. Выборы - форма прямого волеизъявления граждан, осуществляемого в соответствии с Конституцией Российской Федерации, федеральными законами; конституциями (уставами), законами субъектов Российской Федерации; уставами муниципальных образований в целях формирования органа государственной власти, органа местного самоуправления или наделения полномочиями должностного лица.

Итоги выборов позволяют не только выявить расстановку политических сил, но и обнаружить контрасты электоральных предпочтений граждан, отражающие наличие социальных расколов между территориальными общностями.

Особенности электорального поведения и политической идентичности демонстрируют уровень доверия к власти. Также электоральное поведение можно рассматривать в качестве регулятора политической деятельности. Поэтому выявление его особенностей позволяет оценить общественные риски и сделать прогнозы относительно будущей социальной и электоральной ситуации.

В настоящее время выборы главы государства, а также выборы в Государственную Думу выступают в качестве одного из основных событий в жизни современного государства, одной из узловых точек политического процесса. Это особенно верно для президентских режимов. Президентская предвыборная кампания – это фактически центральное событие политического процесса, поскольку в ходе кампании формально определяется и легитимируется главное лицо, принимающее решения, равно как и формулируется повестка дня на срок его полномочий. В следствии этого более подробное изучения состава электората вызывает высокий интерес не только среди исследователей, но и со стороны кандидатов на выборах.

В век информационных технологий ни одно общественное крупное событие не проходит без сопровождения информационной системы, которая в свою очередь порождает массивы информации для дальнейшего анализа. На

основе анализа данных выборов того или иного года можно делать выводы о текущей политической ситуации в стране или конкретном регионе.

Степень научной разработанности данной проблемы. Одной из наиболее актуальных и практически востребованных задач анализа данных является задача разбиения объектов на сравнительно-однородные группы (подмножества), называемые кластерами. Основой для работы иерархического метода кластеризации является так называемая матрица схожести. Быстрый рост объемов обрабатываемой информации, наблюдаемый в последнее время, увеличение размерности решаемых задач обусловливают актуальность разработки и применения методов снижения размерности. Одним из подходов к снижению размерности данных является их кластеризация, то есть объединение в максимально однородные группы.

Целью исследования является сравнение распределения голосов в различных субъектах Российской Федерации по времени.

Задачи исследования:

1) Изучить основные методы кластерного анализа данных.
2) Дать определение иерархической кластеризации данных.
3) Рассмотреть методы иерархической кластеризации данных.
4) Изучить отечественные источники на предмет кластеризации электоральных данных.
5) На основе данных о результатах выборов построить дендрограммы с распределением голосов по годам.
6) Сравнить темпоральную составляющую результатов парламентских выборов в Российской Федерации.

Объект данной работы – результаты президентских и парламентских выборов в Российской Федерации.

Предмет исследования - иерархическая кластеризация электоральных данных.

Теоретическая значимость исследования заключается в возможности использования методов иерархической кластеризации для обеспечения

достаточного уровня наглядности при изображении взаимных связей между объектами из заданного множества.

Структура выпускной квалификационной работы состоит из введения, двух глав, заключения, списка использованных источников и приложения.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе «Задача кластеризации данных» разъясняется понятие «кластеризации данных», рассматриваются этапы применения кластерного анализа, основные метрики для вычисления близости объектов, классификация методов кластеризации, а также методы иерархической кластеризации.

Кластерный анализ (англ. cluster analysis) - многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы[1].

«Кластеризация - это автоматическое разбиение элементов некоторого множества на группы в зависимости от их схожести. Элементами множества могут быть что любые объекты, например, данные или вектора характеристик. Сами же группы принято также называть кластерами»[2].

Существует множество практических применений кластеризации как в информатике, так и в других областях:

1) Анализ данных
   - Упрощение работы с информацией
   - Визуализация данных
2) Извлечение и поиск информации
   - Построение удобных классификаторов
3) Группировка и распознавание объектов
   - Распознавание образов

[1] Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с

[2] Котов, А. Кластеризация данных [Электронный ресурс]. – Режим доступа: http://logic.pdmi.ras.ru/~yura/internet/02ia-seminar-note.pdf

- Группировка объектов

Применение кластерного анализа в общем виде сводится к следующим этапам[1]:

- Отбор выборки для кластеризации.
- Определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства.
- Вычисление значений той или иной меры сходства (или различия) между объектами.
- Применение метода кластерного анализа для создания групп сходных объектов.
- Проверка достоверности результатов кластерного решения.

Проанализировав результаты кластеризации можно скорректировать выбранные параметры, метрику или метод кластеризации, для улучшения результатов.

После выявления вектора характеристик выбирают функцию для определения степени сходства двух объектов, называемую мерой расстояний.

Существуют различные метрики для вычисления близости объектов[2]:

1) Евклидово расстояние. Наиболее распространенная функция расстояния. Является геометрическим расстоянием в многомерном пространстве:

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i')^2}$$

2) Квадрат евклидова расстояния. Применяется для придания большего веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом:

[1] Ершов, К. С. Анализ и классификация алгоритмов кластеризации / К. С. Ершов, Т. Н. Романова. // Новые информационные технологии в автоматизированных системах. – 2016. – №19. – С. 274-279.

[2] https://habr.com/ru/post/101338/

$$\rho(x,x') = \sum_{i}^{n} (x_i - x_i')^2$$

3) Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (т.к. они не возводятся в квадрат). Формула для расчета манхэттенского расстояния:

$$\rho(x,x') = \sum_{i}^{n} | x_i - x_i' |$$

4) Расстояние Чебышева. Это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате. Расстояние Чебышева вычисляется по формуле:

$$\rho(x,x') = \max(| x_i - x_i' |)$$

5) Степенное расстояние. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Степенное расстояние вычисляется по следующей формуле:

$$\rho(x,x') = \sqrt[r]{\sum_{i}^{n} (x_i - x_i')^p},$$

где r и p – параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра – r и p - равны двум, то это расстояние совпадает с расстоянием Евклида.

Выбор метрики зависит от исследователя и решаемой задачи, т.к. результат кластеризации может сильно меняться в зависимости от выбранного варианта.

Общепринятой классификации методов кластеризации не существует, но можно выделить ряд групп подходов[1]:

1) Вероятностный подход. Каждый рассматриваемый объект относится к одному из k классов. Сюда можно отнести: Метод K-средних, K-медиан, EM-алгоритм, Алгоритмы семейства FOREL; Дискриминантный анализ;

2) Подходы на основе систем искусственного интеллекта: Метод нечеткой кластеризации C-средних, Нейронная сеть Кохонена, Генетический алгоритм;

3) Логический подход - построение дендрограммы с помощью дерева решений;

4) Теоретико-графовый подход;

5) Иерархический подход.

Иерархическая кластеризация - совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) вложенных кластеров. Выделяют два класса методов иерархической кластеризации[2]:

Агломеративные методы: новые кластеры создаются путем объединения более мелких кластеров и, таким образом, дерево создается от листьев к стволу;

Дивизивные или дивизионные методы: новые кластеры создаются путем деления более крупных кластеров на более мелкие и, таким образом, дерево создается от ствола к листьям.

Результатом работы алгоритма иерархической кластеризации является дерево разбиений, называемое дендрограммой. Дендрограмма позволяет изобразить взаимные связи между объектами из заданного множества.

---

[1] Бериков В. С., Лбов Г. С. Современные тенденции в кластерном анализе Архивная копия на Wayback Machine // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. - 26 с

[2] Мандель, И. Д. Кластерный анализ / И. Д. Мандель. – М.: Финансы и статистика, 1988. – 176 с

Для построения матрицы сходства (различия) необходимо задать меру расстояния между двумя кластерами. Наиболее часто используются следующие методы определения расстояния[1]:

1) Метод одиночной связи («метод ближайшего соседа»). Расстояние между двумя кластерами полагается равным минимальному расстоянию между двумя элементами из разных кластеров.

2) Метод полной связи («метод дальнего соседа»). Расстояние между двумя кластерами полагается равным максимальному расстоянию между двумя элементами из разных кластеров.

3) Метод средней связи:

Невзвешенный (англ. UPGMA). Расстояние между двумя кластерами полагается равным среднему расстоянию между элементами этих кластеров.

Взвешенный (англ. WPGMA).

4) Центроидный метод.

5) Метод Уорда (англ. Ward’s method). Для оценки расстояний между кластерами используются методы дисперсионного анализа. В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения.

Во второй главе «Кластеризация электоральных данных» речь идет о характере электоральных данных, проведена предварительная обработка данных для исследования, проведен анализ результатов парламентских выборов Российской Федерации за 2011, 2016 и 2021 год, проведено сравнение темпоральной составляющей электоральных предпочтений граждан Российской Федерации.

Данные о результатах парламентских выборов имеют в основном 2 типа источников:

1) Основной - представлен на портале ЦИК и содержит распределение голосов по каждому избирательному участку[1]. И используется при подсчете голосов.

[1] Классификация и кластер. Под ред. Дж. Вэн Райзина. М.: Мир, 1980. 390 с

2) Exit-poll - используемая в мировой социологической практике процедура опроса граждан, производимого социологическими службами на выходе из избирательных участков после голосования. Основными задачами экзитпола являются получение возможности оперативных прогнозов исхода выборов и накопление статистических данных об электорате[2].

Оба типа источников полезны социологам при изучении электоральных предпочтений.

При попытке собрать данные в автоматическом режиме с сайта ЦИК России возникли технические проблемы. Так на сайте используется свой набор шрифтов, который не дает правильно спарсить данные без последующей долгой обработки.

В виду этого для исследования использована уже подготовленная база данных с результатами парламентских и президентских выборов с 2000 по 2021 год, выложенная в общий доступ[3].

Из указанной базы данных для анализа взяты результаты парламентских выборов Российской Федерации за 2011, 2016 и 2021 год, затем проведено сравнение темпоральной составляющей электоральных предпочтений граждан Российской Федерации.

Для удобной работы с данными принято решение собрать всю необходимую нам информацию в 1 xls файл, так данные были загружены на листы с соответствующим названием «2011», «2016», «2021».

После этого было решено убрать из получившегося файла лишние столбцы, а также партии, которые набрали менее 1% голосов в виду ограничения на вычислительные мощности.

---

1 http://www.vybory.izbirkom.ru/region/izbirkom?action=show&global=1&vrn=100100225883172&region=0&prver=0&pronetvd=0

2 http://www.vybory.izbirkom.ru/region/izbirkom?action=show&global=1&vrn=100100225883172&region=0&prver=0&pronetvd=0

3 https://github.com/dkobak/elections/tree/master/data

В итоге для исследования оставлены следующие столбцы в каждом году: Регион; Число избирателей, внесенных в список избирателей на момент окончания голосования; Число действительных избирательных бюллетеней; Политическая партия «Справедливая Россия; Политическая партия «Либерально-демократическая партия России»; Политическая партия «Коммунистическая партия Российской Федерации», Всероссийская политическая партия «Единая Россия». Остальные данные для нашего исследования не потребовались.

Для выполнения анализа было решено использовать язык программирования Python в качестве интерпретатора выступает Anaconda, средой программирования выбран Jupyter Notebook в виду его удобства при использовании. Для решения задачи кластеризации понадобились следующие библиотеки: Pandas, Matplotlib, NumPy, SciPy.

В результате проведенных исследований темпоральная составляющая распределения голосов на парламентских выборах за 2011, 2016 и 2021 год не понесла особых изменений. При явке менее 45% все партии равномерно наращивают голоса, при преодолении порога явки в 45% темп наращивания голосов «Единой России» сильно вырастает, когда как остальные партии сильно замедляются. При переходе порога явки в 60% почти все голоса на избирательных участках начинает получать партия власти. При этом, связи пола избирателей с выбором партии не выявлено, не обнаружено значимой связи и со сферой работы, а также уровнем образования, так как на текущий момент обнаружена связь электоральных данных только с возрастом граждан и географией региона.

Возможно имеет место быть связь «позднего» голосования граждан старше 50 лет с полученными результатами, однако это только гипотеза, так как вопрос повседневных практик и предпочитаемого времени голосования в зависимости от возраста в данном исследовании не поднимался.

ЗАКЛЮЧЕНИЕ

В результате работы была рассмотрена актуальная на сегодняшний день тема иерархической кластеризации электоральных данных Российской Федерации. В процессе исследования описаны основные методы кластерного анализа данных, дано определение иерархической кластеризации данных, рассмотрены методы иерархической кластеризации данных, описаны основные исследования по кластеризации результатов парламентских и президентских выборов в Российской Федерации. На основе данных парламентских выборов за 2011, 2016 и 2021 год проведена кластеризация электоральных данных и построены дендрограммы распределения голосов электората, а также проведено сравнение темпоральной составляющей результатов парламентских выборов в Российской Федерации.

В процессе исследования было отмечено, что распределение электоральной активности на избирательных участках в целом соответствует нормальному распределению. При высоких явках доли голосов за все партии, кроме партии власти, начинают уменьшаться, а все потерянные ими голоса вместе с дополнительными голосами от роста явки переходят к партии власти. Следовательно, можно выдвинуть гипотезу, что граждане, голосующие за партию власти, предпочитают голосовать позже остальных.

Следует отметить, что связи пола избирателей с выбором партии не выявлено, не обнаружено значимой связи и со сферой работы, а также уровнем образования. Однако, удалось обнаружить значимую связь с возрастом респондентов, так граждане старше 51 года чаще выбирают «Единую Россию» и «КПРФ». Интересно и то, что партию власти чаще выбирают в дотационных регионах.

Результаты обработки электоральных данных позволили сделать следующие выводы:

1) Все партии за исключением «Единой России» имеют отрицательную корреляцию с процентом итоговой явки.

2) При явке менее 45% все партии равномерно наращивают голоса, при преодолении порога явки в 45% темп наращивания голосов «Единой России» сильно вырастает, когда как остальные партии сильно замедляются.

3) При переходе порога явки в 60% почти все голоса на избирательных участках начинает получать партия власти.

Можно предположить, что вопрос такого темпа смены политических предпочтений граждан в конце голосования требует дополнительного изучения.