Алгоритм применения метода анализ иерархий совместного с алгоритмом кластеризации

А.Н. Мироненко
mironim84@mail.ru

Омский государственный университет им. Ф.М. Достоевского, Омск, Россия

Аннотация

В работе рассматривается возможность применения известного в математике метода анализа иерархий совместно с алгоритмом кластеризации FOREL для классификации субъектов. Смысл объединения заключается в том, что применяя метод анализа иерархий, а именно принятия решений в условиях определенности, мы подготавливаем данные для дальнейшей работы с ними, а алгоритмом кластеризации (таксономии) происходит их непосредственная обработка. Работу предлагаемого подхода можно условно разделить на два этапа: этап обучения и этап работы. Было проведено компьютерное моделирование проверяющее состоятельность предлагаемого подхода.

Введение

Работа посвящена исследованию возможности применения математических методов в социологических исследованиях. А именно, предлогается алгоритм совместиного применения метода анализа иерархий (МАИ), для предварительной обработки данных которые затем будут кластерезованы.

Вопросу применения математичких методов в других облостях науки посвящено множество публикаций.

В статье [1] автор исследует природу и принципы математического моделирования и его применение в педагогичеких исследованиях.

Кроме того, изучению применения математичского моделирования посвящана книга [2], автор описывает принципы, методы и практики проведения социологических исследований использующие в своей основе математический аппарат.

То есть, можно сделать вывод, что основным и наиболее распространенным математическим методом применяемым в других облостях науки является моделирование.

Мало изученным, в этом плане, является применение хорошо известного в математике метода анализа иерархий.

Его исследованию посвящено достаточно много работ, например, автором статьи [3] рассматриваются области применения МАИ, а так же их особенности. Наиболее важным в применении метода анализа иерархий является правильное построение и нормализация матриц попарных сравнений (pairwise comparison matrices), в статье [4] подробно описывается этот процесс.

In: Sergey V. Belim, Nadezda F. Bogachenko (eds.): Proceedings of the Workshop on Data Analysis and Modelling (DAM 2016), Omsk, Russia, October 2016, published at http://ceur-ws.org

Применение метода анализа иерархий связано с проблемой реверса рангов(rank reversal) — изменения ранжирования альтернатив выбора при их удалении или добавлении, автор работы [5] дает математическое описание данной проблемы и доказывает ее существование.

Важным в данной работе является понятие — кластерный анализ, основные моменты, связанные с ним, были подробно рассмотрены в статье [6]. Автор описывает различные методы кластеризации, кроме того отмечается важная роль выбора расстояния от центра таксона до точек которые будут считаться принадлежащими таксону и координат центра таксонов, показывается, насколько результат кластеризации чувствителен к выбору функции расстояния, использующейся для определения близости точек.

Идея объединения кластеризации и других математических методов рассмотрено в статье [7]. В ней также иследуется возможность совместного применения метода главных компонент, иерархической кластеризации и строгой кластеризации (Principal component methods — hierarchical clustering — partitional clustering) с целью лучшей визуализации данных. Метод главных компонент применяется для предварительной обработки, а методы иерархической и строгой кластеризации для представления данных.

1 Постановка задачи

Поиска новых методов и методологий социологических исследований не вызывает сомнений является актуальной задачей. Наиболее интересным являются вопросы о возможности использования инструментов математического моделирования, а так же информационных технологий для сбора, обработки и последующего анализа данных.

Для социологических исследований важным является обработка данных и дальнейшее их группирование по тем или иным признакам. Еще один важный момент — это поиск определение не типичных объектов, которые нельзя отнести к одной из групп. Эту задачу называют одноклассовой классификацией, выявлением не типичностей, новизны (novelty detection) [8].

С целью решения задачи классификации предметов и обнаружения не типичностей предлагается, применить алгоритм объединения МАИ с одним из методов кластерного анализа и исследовать эффективность данного подхода на практике.

2 Теория

Алгоритма объединения МАИ с одним из методов кластерного анализа состоит из трех шагов:

- Шаг 1. сбор данных;
- Шаг 2. этап подготовки данных (Алгоритм формирования групп);
- Шаг 3. классификации (Алгоритм принадлежности к группе).

Для решения задачи классификации предлагается использовать алгоритм кластерного анализа FOREL. Он работает с точками на n-мерном пространстве, т.е. каждый объект, который необходимо классифицировать представляется в виде точки с n-координатами.

В общем виде схема алгоритма FOREL следующая [9]:

1. Определяем значение T — это минимальный радиус n-сферы на n-мерном пространстве, которая содержит в себе все точки пространства которые кластерезуем.
2. Размещаем центр сферы, с радиусом T, в произвольной точке, что бы выполнялось условие из шага 1.
3. Находим координаты центра сгущения точек, которые оказываются в сфере.
4. Переносим центр сферы в в центр сгущения точек, возвращаемя к шагу 3.
5. Когда центр сферы перестает смещаться считаем, что точки попавшие в нее составляют один класстер.
6. Исключаем их из кластерезуемого множества точек и повтаряем алгоритм начиная с шага 2.

В конечном итоге, после конечного колличества итераций, получаем разбиение исходного множества точек на некоторое колличество кластеров. В каждом из которых расстояния между точками меньше 2T. Однако результат кластеризации, т.е. получаемое колличество кластеров, будет зависит от выбюора начальной точки, в которой находится центр изначальной сферы.

Прежде чем приступить к классификации необходимо подготовить данные для работы с ними. Для этой цели предлагается использовать МАИ.

Работа предлагаемого алгоритма.

Шаг 1. Сбор данных. Выполнение данного шага происходит в виде анкетирования объектов социологического исследования Рис. 1. Объекту прилагается заполнить анкету, в которой ставится задача с определенными критериями выбора и альтернативами ее решения.

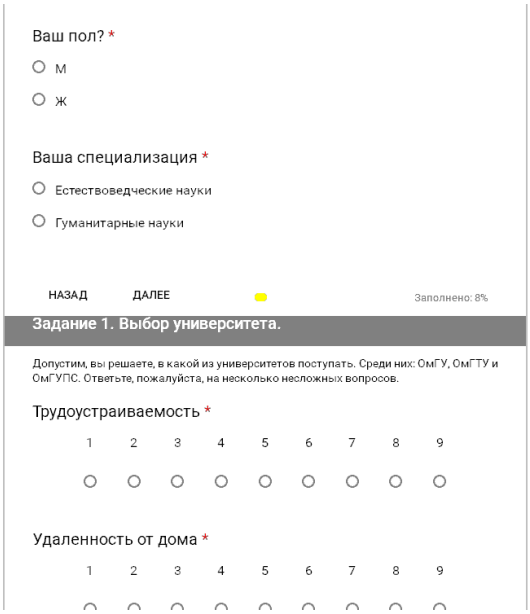

Рис. 1: Привер анкеты для сбора данных

Шаг 2. Подготовка данных.

Алгоритм формирования групп:

1. Объект для каждого из критериев указывает его важность относительно других.
2. Вычисляются относительные веса критериев.
3. Объект указывает насколько каждая из альтернатив предпочтительнее других в пределах каждого критерия.
4. Вычисляются относительные веса альтернативных решений.
5. Вычисляются комбинаторные весовые коэффициенты для каждого из решений.
6. Используя полученные весовые коэффициенты как координаты, получаем точку в n-мерном пространстве.
7. Повторяем шаги с 1 по 6 для всех объектов.
8. Для полученного множества точек, при помощи алгоритма FOREL, решается задача кластеризации.
9. Для каждого таксона определяются координаты центра масс.
10. Таксоны упорядочиваются по величине G=Y/X, где X и Y — координаты центра масс таксона.

Шаг 3. Классификация. После формирования таксонов, проводится анализ каждого из них с целью определить, какую группу объектов он характеризует.

Алгоритм определения принадлежности к группе:

1. Перед новым объектом, который мы хотим классифицировать, ставится задача с определенными критериями выбора и альтернативами ее решения.

2. Объект выполняет шаги 1—6 алгоритма формирования данных для последующей классификации.
3. Определяется принадлежность объекта (n-мерной точки) к одному из таксонов.

3 Результаты эксперимента

С целью апробации предлагаемого алгоритма была написана программа для ЭВМ и проведен эксперимент. Целью эксперимента ставилось определение предрасположенности студента к гуманитарным или техническим наукам (определение технического или гуманитарного склада ума). Всего участвовало в эксперименте 115 студентов различных факультетов ОмГУ им. Ф.М. Достоевского.

В результате выполнения этой стадии эксперимента, была получина следующая кластеризация с различным количеством таксонов (рис. 2).

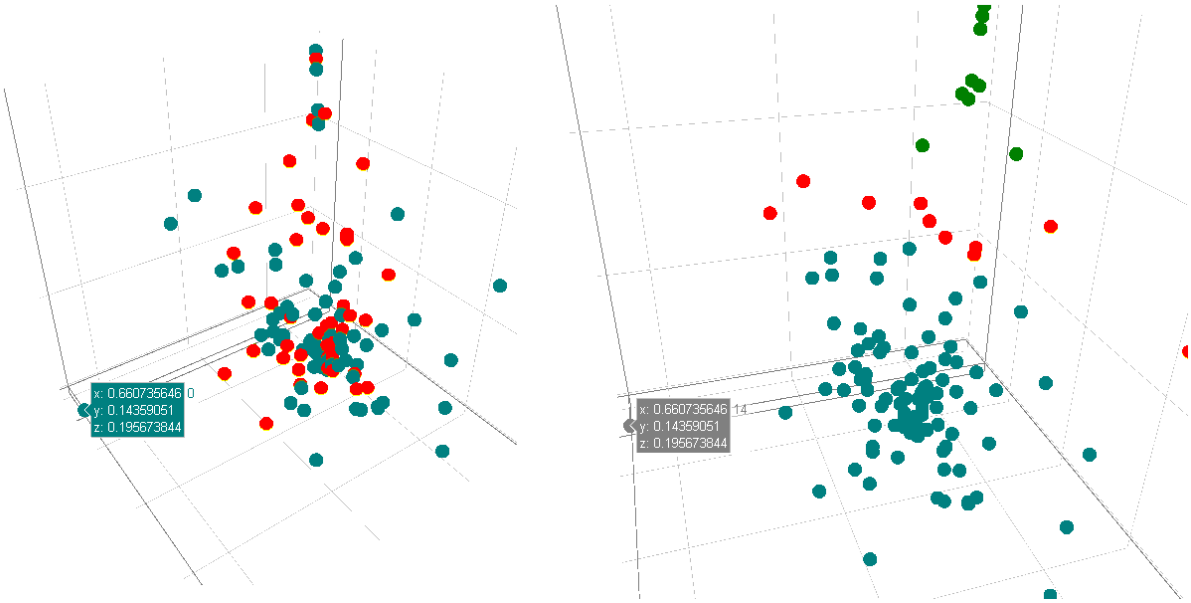

Рис. 2: Пример кластеризации

На Рис. 2 представлено объединение в кластеры с параметром критерия подобия 0.2 — справа, с параметром 0.7 — слева. Поскольку необходимо получить 2 таксона (объекты с техническим и гуманитарного складом ума), экспериментально выбираются параметры, при которых получаем нужное объединение в кластеры. В рамках нашего эксперимента критерий подобия берем равным 0.2. На Рис. 2 таксономия, представленная слева.

Уже на этой стадии проведения эксперимента, можно заметить, что таксоны получаются расположенными очень близко, из-за этого может возникнуть ситуация когда один и тот же объект можно отнести к несколькими таксонам одновременно, очевидно, что такая ситуация не желателен, потому что усложняет процедуру определения принадлежности объекта к группы.

4 Обсуждение результатов

В ходе эксперимента возникли трудности связанные с тем, что при выборе критерия подобия, который даст объединение в кластеры, необходимое для нас, мы получаем очень близко расположенные таксоны, который значительно усложняет классификацию новых объектов. Это может быть связано с неудачным выбором алгоритма объединения в кластеры, поскольку алгоритм FOREL, как правило, применяется в случаях, когда количество таксонов, на которых необходимо выполнить разбиение исходного множества не известен заранее. В нашем эксперименте известно количество таксонов, чтобы получить нужную таксономию необходимо было изменить критерии подобия, что в свою очередь могло отрицательно повлиять на результаты эксперимента.

Тем не менее, изначальное предположение о совместном применение МАИ и алгоритма кластеризации получило свое подтверждение. Таким образом, можно говорить о возможности применения МАИ для подготовки данных для их последующего кластерного анализа.

Выводы и заключение

В работе был предложен алгоритма объединенного применения метода анализа иерархий (принятие решений в условиях определенности) и кластерного анализа. Эксперимент, позволяющий проверять возможность такой объединения, был сделан.

Апробация показала, что применение предложенного алгоритма возможно, но требует дополнительных исследований.

Необходимо подобрать наиболее подходящий алгоритм кластеризации, позволяющий получить определенное количество таксонов и провести новый эксперимент подтверждающий выбор. В случае получения достаточно хорошей таксономии, с четко выраженными областями сгущения точек, можно будет выявлять принадлежность новых объектов исследования к той или иной группе.

В случае, когда количество групп, на которых необходимо выполнить классификацию, неизвестно, использование алгоритма FOREL возможно, но дополнительное исследование каждой из групп необходимо для определения, какие объекты в них входят и в таком случае мы можем решать проблему обнаружения не типичностей, определить объекты, которые нельзя причислить ни к одной группе.

Список литературы

[1] Bakhtiar S. Varaki, Lorna Earl Math Modeling in Educational Research: An Approach to Methodological Fallacies // Australian Journal of Teacher Education, 31(2). 2006. URL: http://dx.doi.org/10.14221/ajte.2006v31n2.3 (access date: 20.10.2016)

[2] Anol Bhattacherjee Social Science Research: Principles, Methods, and Practices // Textbooks Collection. 2012. Book 3. URL: http://scholarcommons.usf.edu/oa_textbooks/3/?utm_source=scholarcommons.usf.edu%2Foa_textbooks%2F3&utm_medium=PDF&utm_campaign=PDFCoverPages (access date: 20.10.2016)

[3] Thomas L. Saaty Decision making with the Analytic Hierarchy Process // International Journal of Services Sciences. 01/2008; є1(01): 83-98; doi: 10.1504/IJSSCI.2008.017590

[4] Andres Farkas The Analysis of the Principal Eigenvector of Pairwise Comparison Matrices // Acta Polytechnica Hungarica. 2007. є4 Issue Number 2; ISSN 1785-8860

[5] Guang Xiao Specification of the AHP hierarchy and rank reversal // Master's Theses. 2010. URL: http://http://udspace.udel.edu/handle/19716/5726?show=full (access date: 20.10.2016)

[6] Cluster Analysis // https://www.qualtrics.com URL: https://www.qualtrics.com/wp-content/uploads/2013/05/Cluster-Analysis.pdf (access date: 10.05.2016).

[7] Francois Husson, Julie Josse, Jerome Pages Principal component methods - hierarchical clustering — partitional clustering: why would we need to choose for visualizing data? // Technical Report Ц Agrocampus. 09/2010

[8] Marco A.F. Pimentel n , David A. Clifton, Lei Clifton, Lionel Tarassenko A review of novelty detection // Signal Processing 99 (2014) 215Ц249. Available online 2 January 2014. ISSN: 0165-1684. URL: https://www.robots.ox.ac.uk/~davidc/pubs/NDreview2014.pdf (access date: 20.10.2016).

[9] Leonid B. Litinskii, Dmitry E. Romanov Neural Network Clustering Based on Distances Between Objects // URL: https://https://arxiv.org/ftp/cs/papers/0608/0608115.pdf (access date: 20.10.2016).

The Algorithm of Applying the Analytic Hierarchy Process in Conjunction with the Clustering Algorithm

Anton N. Mironenko

This paper examines the possibility of applying the analytic hierarchy process, known in mathematics, in conjunction with the FOREL clustering algorithm to classify different subjects. By term "conjunction" we mean a process when the analytic hierarchy process (namely decision making under certainty) is using for preparation of data for further work with them, and the clustering algorithm (taxonomy) is using for direct processing of the data. The proposed approach can be divided into two stages: the training stage and the work stage. We carried out a computer simulation which verifies validity of the proposed approach.