Кластерный анализ многофакторных лингвистических понятий

Соловьев В.Д.

Казанский госуниверситет

solovyev@tatincom.ru

1. Введение

В классической работе Кинэна [1] продемонстрирован сложный характер такого понятия, как “подлежащее”. Им выделено более тридцати простых (элементарных?) признаков, свойственных подлежащим. В различных языках, однако, подлежащие (если вообще имеются) обладают разным набором этих признаков. Таким образом, подлежащее конкретного языка может быть представлено точкой в многомерном пространстве признаков. Такой подход позволяет проводить межъязыковые обобщения исследуемого понятия на основе сопоставления совокупностей признаков, которыми подлежащее обладает в разных языках.

Подобного типа задачи с представлением объектов в многомерном пространстве признаков встречаются в различных областях и в математике уже давно разработаны соответствующие процедуры анализа (называемого кластерным анализом [2]) таким образом представленных данных. Методы

кластерного анализа позволяют выделить множества объектов, близких друг к другу по своим свойствам. Применение кластерного анализа уместно в ситуациях, когда приходится иметь дело с большим числом признаков, так что эксперту становится трудно учесть влияние всех признаков одновременно.

Разработаны пакеты компьютерных программ, реализующих процедуры кластерного анализа. Существующие пакеты обладают большой мощностью и универсализмом, но достаточно сложны для использования. Представляется целесообразным создать для целей лингвистического анализа более простую в эксплуатации компьютерную систему, способную, тем не менее, решать основные задачи кластерного анализа. В работе описывается такая система, а также демонстрируется ее применение на примере анализа системы аффиксов существительных в татарском языке на предмет их “падежеподобности”.

Трудность описания татарских (шире – тюркских) падежей состоит в том, что в татарском языке довольно много аффиксов существительных, причем большинство из них в тех или иных ситуациях имеют словоизменительные функции и могут претендовать на статус падежных. В различных работах по татарскому языку [3, 4] предлагались различные списки падежей. Многофакторный подход к татарским падежам был систематически представлен в [5, 6]. Следует отметить, что до сих пор не существует строгого общепринятого определения падежей.

2. **Общая архитертура и возможности системы**

Разработанная система позволяет решать следующие основные задачи:

А). Разбиение множества заданных объектов на кластеры.

Б). Определение значимости используемых признаков.

В). Поиск и коррекция ошибки в исходных данных.

Для задачи А) входными данными является матрица NxM, где N – число рассматриваемых объектов, M – количество признаков. Общая идея алгоритма кластеризации состоит в следующем. На первом шаге все объекты разбиваются по отдельным кластерам (т.е. каждый кластер будет содержать ровно один элемент). На каждом последующем шаге по определенным формулам вычисляются расстояния между кластерами и два наиболее близких объединяются в один. Процедура объединения кластеров продолжается до тех пор, пока все они не объединятся в один кластер, содержащий все элементы. Процедура кластеризации визуализирована – можно пошагово следить за тем какие кластеры образуются. Конечное решение – сколько должно быть кластеров – принимается пользователем.

Введение в систему задачи Б) связано с тем, что разные признаки могут иметь разную степень важности для характеризации объектов. Хорошо известна [7] идея введения весов признаков, отражающих степень их влияния на кластеризацию объектов. Обычно веса определяются экспертами. В данной системе реализован достаточно редкий подход, связанный с автоматическим определением весов признаков. Веса определяются на основе анализа полученного разбиения. Алгоритм пересчета весов будет описан в следующем разделе.

Система предполагает интерактивное решение задач совместно с пользователем. В частности, предусмотрена возможность введения кластеризации, предполагаемой пользователем, и сравнение ее с построенной системой. Не совпадение этих кластеризаций может вызываться различными причинами, одна из которых – ошибка в исходных данных. Система предоставляет возможность автоматического поиска ошибки. При этом текущая версия системы допускает обнаружение и исправление только единичной ошибки. Идея исправления состоит в том, чтобы выделив те объекты, которые отнесены пользователем и системой к разным кластерам, провести простой перебор всех признаков и последовательно меняя значения признаков выделенных объектов проводить повторную кластеризацию. Процесс продолжается до тех пор, пока не будет обнаружено такое значение одного из признаков, использование которого приводит к кластеризации, указанной

пользователем (в этом случае предполагается, что найдена ошибка) или не будут перебраны все признаки (ошибка не найдена).

3. **Алгоритмы кластеризации и анализа данных**

В данном разделе будут кратко описаны используемые алгоритмы.

Алгоритм кластеризации.

Шаг 0. Вводим кластер с номером i, состоящий из единственного элемента с номером i.

Шаг 1. Формируем матрицу D размера NxN, где N – текущее число кластеров. Элементом D [i, j] является евклидово расстояние между центрами тяжести кластеров i иj.

Шаг 2. Нахождение в матрице D наименьшего элемента (если их несколько, то любого из них). Пусть это будет D [k, l].

Шаг 3. Объединяем кластеры k и l и заменяем кластер k на объединенный. Кластер l исключается из рассмотрения.

Шаг 4. Возврат к шагу 1, пока N (текущее число кластеров) не станет равно 1.

Алгоритм пересчета весов.

Берутся объекты одного из кластеров, и рассматривается первая переменная. Рассчитывается, какое количество объектов принимает то или иное значение по данной переменной. Выбирается то значение, которое принимает наибольшее количество (обозначим его n) объектов из данного кластера. Из числа n вычитается количество всех остальных объектов. Если результат оказался отрицательным, то вес данной переменной для данного

кластера равен 0, иначе полученному положительному значению. Общий вес переменной полагается равным сумме весов переменной по всем кластерам. Таким образом рассчитываются веса для всех переменных.

4. **Кластеризация аффиксов существительных в татарском языке**

В качестве исходных данных были использованы результаты работы [6].

Аффиксы	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Æ	+	+	+	+	+	-	+	+	+	+	+	+		+
-нын	+	+	+	-/+	-/+	-	+	+	+	+	+	+	+	+
-ныкы	+	+	+	+	?	+	+	+	+	+	+	-	+	+
-ны	+	+	+	-	+	+	-	+	+	+	+	+	+	+
-га	+	+	-	+	+	-	+	+	+	+	+	+	+	+
-дан	+	+	-	+	+	-	+	+	+	+	+	+	+	+
-да	+	+	-	+	+	?	+	+	+	+	+	+	+	+
-дагы	+	+	-	+	-	-	+	+	+	+	+	-	+	+
-гы	+	+/-	-	+	-	-	-	-	+/-	-	+	-	+	
-дай	+	+	-	+/-	+	-	+	+	+	+	+	-	-	-
-ча	+/-	+	-	+/-	+	-	+	-	+	+	+	-	-	-

-лы	-	+	-	+	-	-	+	-	+	+	+	-	+	-
-сыз	+/-	+	-	+	+	-	+	(?)	+	+	+	-	+	-

Таблица 1.

В таблице 1 столбы содержат следующие признаки:

1 – Место присоединения аффикса (после аффиксов числа и принадлежности)

2 – Функция (средство связи, не изменяет семантику)

3 – Не выступает, как словообразовательный

4 – Семантика (выражение определенного отношения к управляющему им члену предложения)

5 – Может управляться глаголом

6 – Не может управляться именем

7 – Регулярность присоединения к существительным

8 – Регулярность присоединения к другим классам слов

9 – Присоединение к словосочетаниям

10 – Возможность опущения при однородных членах

11 – Соотносительность с другими падежными аффиксами (синонимия, антонимия и т.д.)

12 – Контекстуальная субстантивация

13 – Ударность

14 – Синтаксические функции (не включают синтаксические функции прилагательного).

Знак + означает, что аффикс обладает соответствующим свойством, знак – означает, что не обладает, остальные знаки соответствуют различным промежуточным случаям.

Предварительная обработка таблицы состояла в замене символов числами. Принята следующая кодировка: + = 4; +/- = 3; отсутствие значения, (?), -(?), +(?) = 2; -/+ = 1; - = 0.

Следует иметь в виду, что в работе [6] были рассмотрены не все аффиксы, которые могут претендовать на статус падежных – за бортом остались аффиксы – лата, -гача, -лык и некоторые другие.

Динамика процесса кластеризации рассматриваемого множества аффиксов видна из следующей таблицы. Столбцы соответствуют последовательным шагам кластеризации, цифра 0 означает, что соответствующий объект остается единственным в своем кластере, другие цифры – это номера кластеров.

Аффиксы	1	2	3	4	5	6	7	8	9	10	11
Æ	0	0	0	0	1	1	3	3	3	3	3

-нын	0	0	0	0	0	3	3	3	3	3	3
-ныкы	0	0	0	0	0	0	0	3	3	3	3
-ны	0	0	0	0	0	3	3	3	3	3	3
-га	1	1	1	1	1	1	3	3	3	3	3
-дан	1	1	1	1	1	1	3	3	3	3	3
-да	0	1	1	1	1	1	3	3	3	3	3
-дагы	0	0	0	0	0	0	0	0	3	3	3
-гы	0	0	0	0	0	0	0	0	0	0	0
-дай	0	0	2	2	2	2	2	2	2	2	3
-ча	0	0	2	2	2	2	2	2	2	2	3
-лы	0	0	0	0	0	0	0	0	0	2	3
-сыз	0	0	0	2	2	2	2	2	2	2	3

Таблица 2.

Анализ полученных данных.

В татарской филологии принято считать падежными аффиксы -нын, -ны,

-га, -дан, -да. Вместе с немаркированным номинативом они образуют классическую шестипадежную систему. Как можно видеть из таблицы 2, на седьмом шаге они выделяются в единый кластер. В работе [8] аффиксы –ныкы и –дагы охарактеризованы, как “падежеподобные”. Они присоединяются к кластеру падежных аффиксов на следующих шагах – 8 и 9. Также следует отметить, что аффикс –гы, реже всего относимый к падежным, присоединяется к кластеру падежных аффиксов в самую последнюю очередь. Полученные результаты свидетельствуют о следующем.

1. Рассматривая задачу о падежности татарских аффиксов как тестовую для созданной системы кластерного анализа, можно констатировать, что система справилась с тестом вполне успешно. Полученная ей классификация татарских аффиксов очень хорошо согласуется с ранее опубликованными в лингвистической литературе результатами.
2. Применение строгих математических методов к весьма полному набору данных позволило подтвердить ранее полученные выводы и повысить степень их надежности. Следует иметь в виду, что во всех предшествующих работах учитывалась лишь часть рассмотренных выше признаков. Кроме того, отнесение аффиксов к падежным или падежеподобным осуществлялось скорее на основе личной лингвистической интуиции исследователей, чем на основе принятой четкой методологии.

5. Заключение

В статье описана компьютерная система кластерного анализа многофакторных лингвистических понятий. Применение методов кластерного анализа позволяет получить объективную классификацию и уместно в тех случаях, когда пространство признаков имеет значительную размерность, так что для эксперта-лингвиста становится трудным обозреть и учесть одновременно все признаки.

Применение разработанной системы к анализу татарских аффиксов

существительных привело к следующим результатам.

Выделение классических падежей (Æ, –нын, -ны, -га, -дан, -да) правомерно, они весьма близки по свойствам и выделяются в кластер. Наиболее близкими к ним являются аффиксы –дагы и –ныкы, присоединяемые к этому кластеру на ближайших шагах. Следующую группу составляют аффиксы –дай, -ча, -лы, -сыз, выделяемые в отдельный кластер. Наиболее далеко от падежных отстоит аффикс –гы.

Литература

Кинэн Э.Л. К универсальному определению подлежащего. В сб. “Новое в зарубежной лингвистике”, вып. 11, М.: Прогресс, 1982.

Everitt B. Cluster analysis. London: Heinemann, 1981.

Закиев М.З. К вопросу о категории падежа в тюркских языках. В сб. “Вопросы тюркологии и истории востоковедения”, Казань: КГУ, 1964.

Ганиев Ф.А. О синтетических и аналитических падежах в татарском языке. Вопросы тюркологии, Казань, 1970.

Ирисов Н.И. Падежное и непадежное в татарском языке. Труды межд. конф. “Языки, духовная культура и история тюрков: традиции и современность”, Казань, 1992.

Сулейманов Д.Ш. К вопросу о числе падежей в татарском языке. В сб. “Исследования в лингвистике”, Казань: Фэн, 1996.

Мандель И.Д. Кластерный анализ. М.: Финансы и статистика, 1988.

Тумашема Д.Г., Ирисов Н.И. К вопросу о падежном характере аффиксов –дагы и –ныкы. “Советская тюркология”, N6, 1989.