УДК 519.254+519.237.8+519.237.7

Визуальные представления в задаче кластеризации корпуса текстов

А.Р. Насибуллина[1]

[1]Московский физико-технический институт (государственный университет)

Когда мы имеем дело с большой совокупностью текстов, нам нужно в первую очередь выявить в них структуру. Инструменты Text Mining дают нам такую возможность: документы преобразуются в числовую форму и затем группируются на основании различных методов кластерного анализа. Применение автоматических методов кластеризации практически полезно, если результат можно проинтерпретировать и обосновать найденные закономерности. Визуальный интерактивный кластерный анализ является одним из эффективных способов, чтобы обеспечить понимание и интерпретируемость результатов. Мы опишем простой программный инструмент для изучения корпуса текстов и их кластеризации [1]. Программа включает в себя 4 метода визуализации на стадии препроцессинга и в ходе последующего отбора объектов.

Входные данные должны быть предварительно параметризованы. Здесь используется традиционная векторная форма документов. Для параметризации выбирается список ключевых слов. Уровень специфичности слова w в корпусе текстов определяется числом $K > 1$, которое показывает во сколько раз частота $f(w)$ этого слова в корпусе текстов превышает его частоту $F(w)$, с которой оно встречается в лексике в целом: $K = f(w)/F(w)$. В работе мы используем программу LexisTerm [2], которая реализует этот критерий специфичности. Список ключевых слов корректируется экспертом, затем все тексты представляются в векторной форме относительно этого списка.

Программа включает 4 метода визуализации: диаграмма межобъектных расстояний, представление в подпространстве параметров объектов, представление в подпространстве пар главных компонент, представление на диаграмме двух альтернатив [3,4]. Первый метод позволяет определить возможное число кластеров в данных. Для этого рассчитываются расстояния между всеми объектами, и строится гистограмма. Число локальных максимумов гистограммы равно $K = n(n+1)/2$, где n - нижняя граница числа кластеров. Второй и третий методы хорошо известны и позволяют представить многомерные объекты в двумерном пространстве пар параметров и главных компонент. Подпространства выбираются пользователем в интерактивном режиме. Четвертый метод позволяет представить многомерные объекты в двумерном пространстве двух альтернатив, выбранных экспертом, и зрительно оценить по принципу соседства близость

остальных объектов к ним. При визуализации также используется метод «меченых атомов», что позволяет классифицировать объекты по принципу соседства.

В качестве данных для эксперимента использовался корпус из 100 текстов, соответствующих тематике детского здравоохранения в Европе. Список ключевых слов включал в себя 10 слов, и каждый текст был представлен в векторной форме относительно этого списка. На диаграмме межобъектных расстояний (рис.1) наблюдается 5 локальных максимумов, следовательно, данные включают как минимум 3 кластера. Данные были представлены графически в подпространствах параметров, главных компонент и двух альтернатив. Пользователь зрительно может выделить эти кластеры в данных, а также на основе принципа соседства определить близость объектов к двум альтернативам и маркированным объектам.

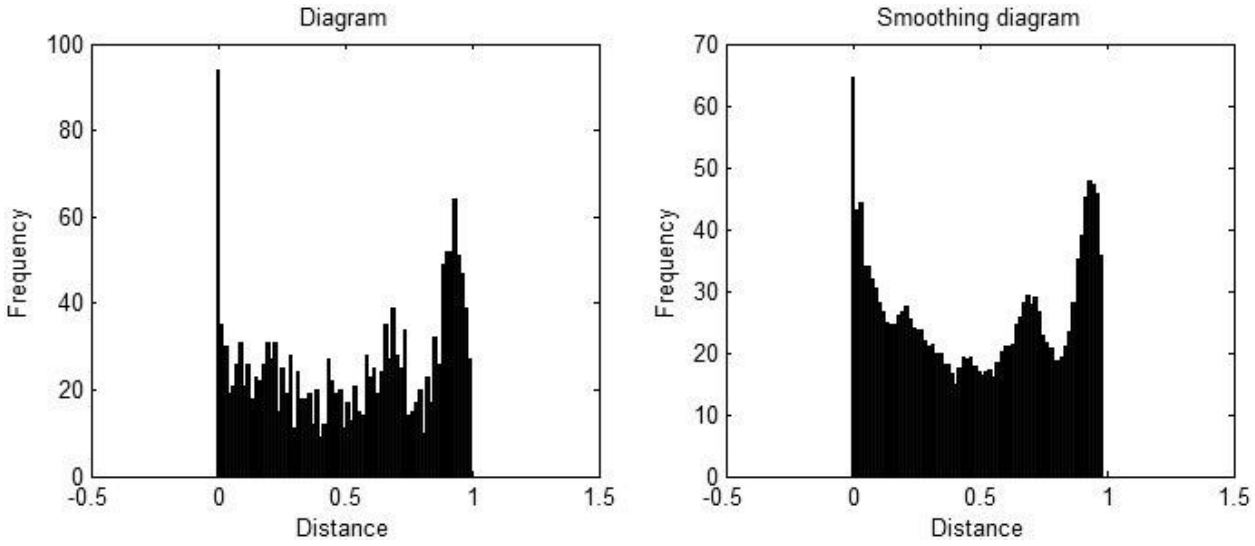

Рис.1 Диаграмма межобъектных расстояний

Литература

[1] *Насибуллина А.* Визуальные представления в задаче отбора многомерных экономических объектов // Труды 57-й научной конференции МФТИ с международным участием, посвященной 120-летию со дня рождения П.Л.Капицы. – 2014.- С.119-121.

[2] *Lopez R., Alexandrov M., Tejada J.* LexisTerm - the program for term selection by the criterion of specificity. - Proc. of 4-th Intern. Conf. On Intelligent Information and Engineering Systems, 2011, vol. 24, ITHEA. - pp. 8-15.

[3*] Насибулина А.* Визуальный кластер-анализ многомерных объектов. - Современная экономика: теория, политика, инновации. - Москва, 2014. - С.79-84.

[4] *Nasibullina A., Alexandrov M., Kovaldji A.* Simple Free-Share package for Visual Analysis of multidimensional data sets. - Computational Models for Business and Engineering Domains, 2014, vol.30, ITHEA. - pp. 216-224.