но считать проекты SearchMonkey от Yahoo, Rich Snippets от Google, или Bing Powerset. Такие системы подтверждают тот факт, что семантический поиск является перспективным направлением развития поиска информации, и решение проблем семантического поиска откроет новые возможности как для отдельно взятых компаний, так и человечества в целом.

Библиографический список

1. Allemang D., Hendler J. Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL // Morgan Kaufmann, 2008.

Bernstein A., Kaufmann E., Fuchs N. Talking to the semantic web – a controlled english query interface for ontologies // AIS SIGSEMIS Bulletin. – 2005. – № 2. – P. 42–47.

2. Corcho O. Ontology based document annotation: trends and open research problems // Int. J. Metadata, Semantics and Ontologies. – 2006. – № 1. – P. 47–57.

3. Хорошевский В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) / В. Ф. Хорошевский // Искусственный интеллект и принятие решений. – 2008. – № 1.

УДК: 519.237.8

Описание одного алгоритма кластеризации типа Forel

В.В. Журавлева, А.А. Бондарева
АлтГУ, г. Барнаул

В научной литературе описано множество методов автоматической классификации. Среди них есть достаточно простые (k-средних, Forel), применение которых позволяет построить классы простой формы, и трудоемкие (KRAB), но строящие при этом «необычные» классы [1, 2]. Здесь описан простой алгоритм, позволяющий получать классы необычной формы.

Предварительно стоится разбиение массива данных (представленного таблицей числовых значений признаков, в которой строки соответствуют объектам, а столбцы признакам) на большое количество мелких непересекающихся кластеров с помощью алгоритма Форел. Также для начального разбиения можно использовать иной алгоритм. В итоге каждый кластер можно описать сферой определенного радиуса. Суть алгоритма состоит в последовательном объединении маленьких сфер-кластеров в большие классы по принципу «ближайшего соседа».

Опишем некоторые понятия.

Центр кластера вычисляется покоординатно по формуле:

$$z_i = \frac{1}{n}\sum_{j=1}^{n} x_{ji},$$

где n – количество объектов в кластере, x_{ji}, z_j – значения i-го признака для j-го объекта и центра кластера соответственно.

Радиус кластера – максимальное расстояние объектов от центра кластера. Радиус рассчитывается по следующей формуле:

$$R = \max_{j=1,..n} \sqrt{\sum_{i=1}^{m} \left(z_i - x_{ji}\right)^2},$$

где m – количество признаков.

Расстояние между кластерами рассчитывается как расстояние между центрами сфер-кластеров, за вычетом радиусов этих сфер. Формула расстояния между кластерами:

$$P_{kt} = \sqrt{\sum_{i=1}^{m} \left(z_{ki} - z_{ti}\right)^2} - \left(R_k + R_t\right),$$

где z_{ki}, z_{ti} – значения i-го признака для центров k-го и t-го кластера, R_k и R_t – радиусы k-го и t-го кластера.

Условие присоединения сфер-кластеров определяется как

$$P_{kt} \le C\ .$$

Параметр C – пороговое значение, задаваемое пользователем. Объединение кластеров происходит до тех пор, пока все оставшиеся расстояния между кластерами не будут больше C.

Итак, на вход алгоритма подается совокупность маленьких сфер-кластеров.

Пошаговое выполнение алгоритма:

1. Найти центр и радиус каждого кластера.
2. Задать пороговое значение (параметр C).
3. Найти матрицу расстояний между кластерами.
4. В порядке возрастания межклассового расстояния объединять кластеры до тех пор, пока оно не станет больше C.
5. Вывести результат классификации.

Пункт 5 предполагает выполнение цикла.

В частных случаях выполнения данного алгоритма могут быть получены цепочки либо деревья из маленьких сфер-кластеров.

Замечание 1: для радиусов кластеров и межклассовых расстояний использована Евклидова метрика, которую можно заменить на любую другую [3].

Замечание 2: допустимо применение описанного алгоритма при начальном разбиении на пересекающиеся классы, в этом случае алгоритм работает корректно.

Замечание 3: последовательно увеличивая значение параметра C, можно получать меньшее количество классов. Таким образом, данный алгоритм можно отнести к иерархическим методам классификации. Алгоритм легко модифицировать в вариант, при котором количество классов не будет превышать максимально допустимое.

Корректность алгоритма проверена на модельных примерах, а также по совокупности данных о вызовах скорой помощи (по сердечно-сосудистым заболеваниям) и ряду геофизических факторов за 2006-2010 г.г. Для указанных данных ранее проводились исследования по малому массиву данных с применением простых методов кластеризации [4, 5]. С использованием описанного алгоритма построена кластерная структура, которая позволяет обнаружить значимую эмпирическую зависимость между количеством вызовов скорой помощи и значениями геофизических факторов.

Библиографический список

1. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999. – 270 с.

2. Половикова О.Н. Применение кластерного анализа для математического моделирования информационно-поисковых систем. [Электронный ресурс] Режим доступа: http://www.ict.nsc.ru/ws/YM2003/6285/.

3. Половикова О.Н., Фокина В.В. Использование евклидова и манхэттенского расстояний в качестве меры близости для решения задачи классификации // Известия Алтайского государственного университета. – Барнаул, 2010. – № 1-1(65). – С. 101-102.

4. Журавлева В.В. Исследование связи между состоянием геомагнитного поля и обострением сердечно-сосудистых заболеваний // Известия Алтайского государственного университета. – Барнаул, 2011. – №1-1(69). – С. 98-100.

5. Журавлева В.В., Егошин А.В. Применение кластерного анализа для обнаружения влияния ГМП на обострение сердечно-сосудистых заболеваний // МАК-2010: материалы тринадцатой региональной конференции по математике. – Барнаул: Изд-во Алт. ун-та, 2010. – С.86-87.