УДК 519.7 : 681.3

О.Н. Половикова, В.В. Фокина

Использование евклидова и манхэттенского расстояний в качестве меры близости для решения задачи классификации

Ключевые слова: кластеризация, метрика, функция близости, критерий выбора, экспертная система.
Key words: clusterization, metrics, proximity function, selection criterion, expert system.

Рассмотрим следующую задачу классификации. Пусть в n-мерном признаковом пространстве расположено k-объектов из класса A. Объект класса B также характеризуется n-признаками, но в общем случае значение каких-то признаков может быть не определено (не задано) или для одного признака может быть задано несколько значений. Кроме этого, следует учитывать актуальность (необходимость) каждого признака. Например, признак (характеристика) объекта *B* принимает значение *X1* и *X2*, а актуальность (важность) данного признака может быть определена значением шкалы «важная характеристика; скорее важная характеристика, чем нет; характеристику можно не учитывать».

Для объекта класса B необходимо определить наиболее близкий, с точки зрения совпадения значений по n-признакам, объект класса *A*. Искомая функции близости между объектами классов *B* и *A* должна учитывать многозначность в определении некоторых признаков, а также актуальность каждого признака.

Рассматриваемая задача нахождения наиболее близкого объекта из некоторого класса к определенному заданному объекту другого класса является востребованной для многих прикладных областей. В рамках данного исследования рассмотрим один из способов решения задачи определения научного руководителя для студента. Концепция предлагаемого метода решения строится на поиске наиболее близкого объекта к объекту другого класса.

Чтобы выбрать научного руководителя из определенной группы преподавателей кафедры или факультета для конкретного студента, необходимо сформировать критерий выбора. Критерий выбора строится по характеристикам преподавателя с указанием их значений, перечень необходимых характеристик каждый конкретный студент определяет самостоятельно. Другими словами, для формирования критерия выбора студент должен по шаблону оформить свои предпочтения относительно будущего научного руководителя. Из предпочтений студента формируется объект класса *Образ* (с описанием значений признаков и их актуальностью). Задача поиска научного руководителя сводится к задаче определения наиболее близкого экземпляра класса *Преподаватель* (класс потенциальных научных руководителей) для сформированного объекта класса *Образ*.

Если каждый экземпляр класса *A* рассматривать как отдельный кластер, тогда можно говорить о классической задаче кластеризации, где заданный объект класса *B* требуется классифицировать (определить) в один из кластеров исходя из некоторой меры близости. В кластерном анализе для количественной оценки близости вводится понятие *метрики.* Сходство и различие между классифицируемыми объектами устанавливаются в зависимости от метрического расстояния между ними. Рассматриваемая задача классификации сводится к задаче определения функции близости между объектами классов *B* и *A* – выбора меры расстояния между объектами.

В кластерном анализе используют различные меры расстояния между объектами [1, с. 156–160]:

Евклидово расстояние – наиболее общий тип расстояния (см. формулу 1).

Расстояние «городских кварталов» – по сравнению с евклидовым расстоянием влияние отдельных больших разностей (выбросов) уменьшается, так как они не возводятся в квадрат (см. формулу 2).

Расстояние Махаланобиса применяется в случае ненулевой корреляции переменных. В рассматриваемой задаче поиска научного руководителя корреляция между значениями признаков объектов равна нулю, поэтому расстояние Махаланобиса эквивалентно квадратичному евклидову расстоянию.

На практике используются и другие метрики, но большинство из них являются частными формами специального класса метрических расстояний, известных как метрики Минковского.

Учитывая специфику решаемой задачи, остановимся на использовании евклидова и манхэттенского расстояний в качестве меры близости между экземплярами класса *Преподаватель* (класс потенциальных научных руководителей) и класса *Образ* (объект, описывающий предпочтения конкретного студента).

Евклидово расстояние с весовыми коэффициентами признаков:

$$P_{ev}(X_i , X_j) = \sum_m[(1/w_m)*(x_{im} - x_{jm})]^2, \qquad (1)$$

где X_i, X_j – координаты i-го и j-го объектов в n-мерном пространстве; x_{im}, x_{jm} – величина m-той характеристики у i-го (j-го) объекта (m = 1,2,...,n; i, j = 1,2,...,t); t – количество объектов; w_m – весовой коэффициент m-го признака.

Расстояние city-block (городских кварталов) или манхэттенское расстояние с весовыми коэффициентами признаков:

$$\rho_{cb}(X_i, X_j) = \sum_m [(1/w_m)| x_{im} - x_{jm} |], \quad (2)$$

где X_i, X_j – координаты i-го и j-го объектов в n-мерном пространстве; x_{im}, x_{jm} – величина m-той характеристики i-го (j-го) объекта (m = 1,2,...,n; i, j = 1,2,...,t); t – количество объектов; w_m – весовой коэффициент m-го признака.

Расстояние между двумя объектами складывается из суммы разностей значений признака (квадрата разностей или абсолютных разностей) двух объектов. С учетом специфики рассматриваемой задачи каждое слагаемое суммы следует разделить на весовой коэффициент w_m, отражающий актуальность соответствующего признака.

Так как один признак может характеризоваться несколькими значениями (множественность признака больше 1), тогда каждое слагаемое может быть представлено в виде суммы разностей нескольких значений одного признака или каждое значение (каждая характеристика) одного признака может быть рассмотрено как значение отдельного признака. В том и другом случае весовой коэффициент, отражающий актуальность соответствующего признака, может быть установлен одинаковым для всех значений признака или индивидуально выбран для каждого значения.

В ходе проведенного исследования (в том числе и по результатам анкетирования) определены и нормированы основные характеристики преподавателей, значения которых для объекта класса *Преподаватель* выступают в качестве координат n-мерного пространства. Среди основных значимых характеристик можно выделить следующие признаки: научное направление, в котором специализируется преподаватель (данный признак может принимать несколько значений); ученая степень преподавателя; стаж работы преподавателя в данном направлении; количество публикаций преподавателя по данному направлению; количество выпускников.

Для реализации критерия выбора научного руководителя построена экспертная система [2]. База данных экспертной системы содержит характеристики преподавателей одной из кафедр социологического факультета АлтГУ. Дерево решений представляет собой набор правил для вычисления функции близости между объектами, представленными в базе данных, и объектом класса *Образ*, характеристики которого определяет студент (пользователь экспертной системы). Найденные значения функции близости задают *рейтинг* (предпочтение выбора данного преподавателя в качестве научного руководителя для пользователя) для каждого объекта класса *Преподаватель*.

Каждое решение экспертной системы состоит из списка объектов класса *Преподаватель* с указанием их характеристик, а также рейтинга данного объекта (два значения). Рейтинг объекта вычисляется по формулам евклидова и манхэттенского расстояний между этим объектом и объектом класса *Образ* (см. формулы 1 и 2).

Реализация экспертной системы выполнена с использованием языка PHP и СУБД MySQL. Апробация разработанной системы показала необходимость использования данной системы для организации учебного процесса.

Библиографический список

1.Факторный, дискриминантный и кластерный анализы: пер. с англ. Дж.-О. Ким, Ч.У. Мьюллер, У.Р. Клекка и др.; под. ред. И.С. Енюкова. – М., 1989.

2. Статические и динамические экспертные системы : учеб. пособие / Э.В. Попов, И.Б. Фоминых, Е.Б. Кисель, М.Д. Шапот. – М., 1996.