*А. А. Горбачев, И. А. Жалнин, К. А. Ищенко**

К ВОПРОСУ СРАВНИТЕЛЬНОЙ ЭФФЕКТИВНОСТИ МЕТОДОВ СРАВНЕНИЯ ОБРАЗОВ

Для двух заданных образов требуется построить оценку тематической близости $rank(a,b)$ так, что при $rank(a,b)=1$ тематики прообразов a и b полностью совпадали и для того, чтобы выполнялось $rank(a,b)>rank(a,c)$, было необходимо и достаточно того, что прообраз b был ближе по тематике к прообразу a, чем прообраз c.

Эффективность конечного метода решения той или иной задачи в значительной степени зависит от эффективности используемой схемы ранжирования, от тестовых данных. Рассмотрим оценку средней эффективности, т.е. эффективности при условии случайных данных. Для этого строим метрику на множестве схем ранжирования так, что по любым двум известным эффективностям и расстояниям между ними исследуемым методом указанную оценку можно построить по формуле

$$ef_c = \begin{cases} ef_a - \|a-c\|, & \|a-c\|+\|b-c\|>\|a-b\| \wedge \|a-c\| \le \|b-c\| \\ ef_a + \|a-c\|, \end{cases} \quad (1)$$

где ef_a – эффективность метода, и известно, что эффективность метода b выше. В качестве метрики – сравнительная эффективность двух методов – модуль разности средних эффективностей методов.

В частности, для задачи поиска

$$e(f) = \int_0^1 p_f dr_f. \quad (2)$$

Таким образом, конечная форма сравнительной эффективности

$$\Delta(f,g) = \left| \int_0^1 p_f dr_f - \int_0^1 p_g dr_g \right|. \quad (3)$$

Для построения оценки сравнительной эффективности во всем классе схем ранжирования используем алгоритм:

1. Выделение всех свойств и построение общего словаря T.
2. Преобразование $D \to R^{|T|}$ по формуле

---

* Работа выполнена под руководством канд. техн. наук, доцента ФГБОУ ВО «ТГТУ» М. А. Ивановского.

$$\vec{d}_i = (d_{i1}, d_{i2}, \ldots, d_{i|T|});$$
$$d_{ij} = \begin{cases} f(d_i, t_k), t_k \in d_i; \\ 0, t_k \notin d_i. \end{cases} \quad (4)$$

Сравнение полученных образов и вычисление ранга

$$rank(d_1, d_2) = \vec{d}_1 X \vec{d}_2, \quad (5)$$

где X – матрица размерности $|T| \times |T|$. Функции f могут различаться для первого и второго аргументов. Далее обозначим f_1 – функцию первого аргумента, f_2 – второго. Эти функции выбираются таким образом, чтобы отражать встречаемость и значимость терма в той или иной структуре, поэтому будем называть их функциями значимости.

Сравнительная эффективность методов сравнения образов коррелирует с вероятностью совпадений знаков сравнительной оценки близости соответствующих прообразов по всей совместной базе.

$$\rho(m_1, m_2) = P\{(rank_{m_1}(a,b) - rank_{m_1}(a,c))(rank_{m_2}(a,b) - rank_{m_2}(a,c)) > 0. \quad (6)$$

Подставим в форму (6) вид рассматриваемых схем (5):

$$\rho(m_1, m_2) = P\{(\vec{a}^T X_1 \vec{b} - \vec{a}^T X_1 \vec{c})(\vec{a}^T X_2 \vec{b} - \vec{a}^T X_2 \vec{c}) > 0\} \Leftrightarrow$$
$$\rho(m_1, m_2) = P\{\vec{a}^T X_1 (\vec{b} - \vec{c}) \vec{a}^T X_2 (\vec{b} - \vec{c}) > 0\}. \quad (7)$$

Проведя замену переменных $\vec{x} = \frac{\vec{a}}{\|\vec{a}\|}$, $\vec{y} = \frac{(\vec{b} - \vec{c})}{\|\vec{b} - \vec{c}\|}$ и учитывая то, что $\|\vec{b} - \vec{c}\|\|\vec{a}\|$ не влияют на знак рассматриваемого выражения, получим

$$\rho(m_1, m_2) = P\{\vec{x}^T X_1 \vec{y} \vec{x}^T X_2 \vec{y} > 0\}. \quad (8)$$

В том случае, когда функции значимости равны и симметричны относительно нуля, значения *x* и *y* можно рассмотреть как независимые случайные величины, равномерно распределенные по единичной сфере с центром в $\vec{0}$. Далее можно разделить указанное событие на два непересекающихся:

$$\rho(m_1, m_2) = P\{\vec{x}^T X_1 \vec{y} > 0 \wedge \vec{x}^T X_2 \vec{y} > 0\} + P\{\vec{x}^T X_1 \vec{y} < 0 \wedge \vec{x}^T X_2 \vec{y} < 0\}. \quad (9)$$

Рассмотрен только тот случай, когда документы *b* и *c* отличаются ровно на один терм, причем этот терм равномерно распределен по словарю. Таким образом переходим к системе

$$\rho(m_1,m_2)=\frac{1}{|T|}\sum_{i=0}^{|T|}\left(P\{\vec{x}^T\vec{x}_1^i>0\wedge\vec{x}^T\vec{x}_2^i>0\}+P\{\vec{x}^T\vec{x}_1^i<0\wedge\vec{x}^T\vec{x}_2^i<0\}\right);$$
$$\rho(m_1,m_2)=\frac{1}{|T|}\sum_{i=0}^{|T|}\left(P\{\vec{x}_1^{iT}\vec{x}>0\wedge\vec{x}_2^{iT}\vec{x}>0\}+P\{\vec{x}_1^{iT}\vec{x}<0\wedge\vec{x}_2^{iT}\vec{x}<0\}\right). \quad (10)$$

Рассмотрим сечение пространства плоскостью $\left(\vec{x}_1^i,\vec{x}_2^i\right)$, проходящей через начало координат. Плоскость будет ортогональна пересечению гиперплоскостей $\vec{x}_1^{iT}\vec{x}=0$ и $\vec{x}_2^{iT}\vec{x}>0$. Сечение единичной сферы представляет собой окружность, из чего следует, что вероятность $P\{\vec{x}_1^{iT}\vec{x}>0\wedge\vec{x}_2^{iT}\vec{x}>0\}+P\{\vec{x}_1^{iT}\vec{x}<0\wedge\vec{x}_2^{iT}\vec{x}<0\}$ для векторов из выбранного сечения составляет $\frac{1}{\Pi}\left(\Pi-\arccos\left(\frac{\left(\vec{x}_1^i,\vec{x}_2^i\right)}{\left\|\vec{x}_1^i\right\|\cdot\left\|\vec{x}_2^i\right\|}\right)\right)$. Для любого движения вдоль вектора $\vec{r}\in\left\{x\middle|\vec{x}_1^i\vec{x}=0\cap\vec{x}_2^i\vec{x}=0\right\}$ эта ситуация будет сохраняться, из чего можно сделать вывод о том, что полученная формула верна для любых векторов из единичной сферы.

Очевидно, что одно и то же множество объектов можно разбить на кластеры различными способами, или при использовании одного метода можно получить целую группу различных разбиений. В таком случае имеет смысл определить качество разбиений с целью выбора наилучшего разбиения, т.е. сформулировать количественный критерий, в соответствии с которым можно было бы предпочесть одно разбиение другому. Для формулировки представлений о качестве классификации в постановку задачи вводится функционал качества разбиения, задающий способ сопоставления с каждым разбиением P числа $Q(P)$, которое оценивает в некоторой шкале степень оптимальности разбиения P. То разбиение P^*, на котором выбранный функционал достигает экстремального значения, считается наиболее предпочтительным. Данное направление в кластерном анализе называется оптимизационным и вводит задачу классификации в сугубо математическое русло, так что в математической постановке задача классификации сводится к поиску оптимального разбиения P^* и формулируется в виде

$$Q(P)\to\underset{P\in\Pi}{\text{extr}}, \quad (11)$$

где Π – множество всех возможных разбиений исходного множества объектов X. Математические методы и реализующие их кластер-процедуры, доставляющие экстремум выбранному функционалу, соответственно называют оптимизационными.

В продолжение рассмотрения оптимизационных методов решения задачи классификации следует указать на то обстоятельство, что, как отмечал С. А. Айвазян, «в статистической практике выбор функционала качества разбиения $Q(P)$ обычно осуществляется весьма произвольно, опирается скорее на эмпирические и профессионально-интуитивные соображения, чем на какую-либо точную формализованную схему... Однако ряд распространенных в статистической практике функционалов качества удается постфактум обосновать и осмыслить в рамках строгих математических моделей» [2].

Если классификация, которую требуется найти, описывается матрицей определенной структуры, к примеру, матрицей отношения эквивалентности, то задача заключается в оценке параметров искомой структуры так, чтобы искомая структура минимально отличалась бы от исходной структуры. Иными словами, отношение, отвечающее исходным данным, необходимо аппроксимировать отношением, которое отвечает представлению о наилучшей классификации, так что это направление решения задач кластер-анализа именуется аппроксимационным; в этом случае проблема сводится к следующей оптимизационной задаче:

$$\|M\| \to \text{extr}, \tag{12}$$

где $M = J - BX$. В этом равенстве J обозначает отношение, которое требуется найти, X – матрица исходных данных, B – оператор перехода от X к J, а $\|M\|$ обозначает некоторую норму. На практике применяют методы аппроксимации как матрицы «объект–свойство», так и матрицы «объект–объект».

Список литературы

1 Добрынин, В. Ю. Автоматическая классификация документов / В. Ю. Добрынин, И. Е. Кураленок // Интернет и современное сообщество : тр. Всерос. науч.-метод. конф. – СПб., December 1998.

2 Кураленок, И. Е. Автоматическая классификация документов с использованием семантического анализа / И. Е. Кураленок, И. С. Некрестьянов // Электронные библиотеки : тр. Первой Всерос. науч. конф. – СПб., октябрь 1999.

Кафедра «Информационные системы и защита информации»
ФГБОУ ВО «ТГТУ»