Результаты компонентного и кластерного анализов совпали.

На основании выявленного списка компетенций и критериев их оценки была сформирована карта компетенций системного аналитика различный уровней (рисунок 3), которая отражает требуемый уровень знаний на указанных должностных позициях.

Таким образом, карта компетенций показывает характеристику сотрудника, его способности к выполнению тех или иных трудовых функций.

Литература

1. Гузаиров М.Б. Информационное, математическое и программное обеспечение поддержки принятия решений при отборе претендентов / Гузаиров М.Б., Сметанина О.Н., Сафиуллина Д.Ф., Маркушева А.М. // Вестник уфимского государственного авиационного университета. – 2014. – №5(66). – С. 185-191.

2. Ильясов Б.Г. Интеллектуальная информационная система поддержки процедур управления воспроизводственным процессом / Ильясов Б.Г., Макарова Е.А., Павлова А.Н. // Программные продукты и системы. – 2010. – №1. – С. 88-90.

3. Тархов С.В. Информационная система отбора претендентов на вакантные рабочие места / Тархов С.В., Минасова Н.С., Шагиева Ю.Р. // Вестник башкирского государственного аграрного университета. – 2012. – №4(24). – С. 88-92.

Н.Г. Крупец, С.В. Федоров

МЕТОДЫ ОБРАБОТКИ РАЗНОТИПНЫХ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

(Самарский национальный исследовательский университет им. академика С.П. Королева)

Введение

В представленной работе рассматриваются методы построения логических решающих правил классификации объектов, характеризуемых вектором признаков, измеренных в шкале наименований.

Постановка задачи

Разработать систему – сайт для исследования методов построения классификаторов в пространстве разнотипных признаков. Под разнотипным признаками понимают данные, измеренные в разных шкалах, например, в шкале наименований.

Шкала наименований это - качественная шкала, она не содержит количественную информацию, в ней нет нуля и единиц измерений. Элементы этих шкал характеризуются только соотношениями эквивалентности (равенства) и сходства конкретных качественных проявлений свойств. Примером

может служить атлас цветов (шкала цветов). Процесс измерения заключается в визуальном сравнении окрашенного предмета с образцами цветов (эталонными образцами атласа) [1].

Методы построения классификаторов

Существуют разные методы построения классификаторов [2]:

- метод, основанный на алгоритме DW (реализует направленную процедуру поиска в виде дерева и дает локально-оптимальное решение);
- метод, основанный на алгоритме ND (реализует решающее правило на основе набора деревьев);
- метод, основанный на алгоритме CORAL (реализует направленную процедуру поиска закономерностей).

Остановимся на методе построения классификаторов на основе алгоритма CORAL, так как он обладает наибольшим быстродействием, что очень важно при работе с выборкой большого объема.

Этот алгоритм реализует направленную процедуру поиска закономерностей, характеризующих образ w, и дает локально-оптимальное решение. Работа алгоритма состоит из нескольких этапов.

Этап 1. В начале поиска путем перебора всевозможных элементарных высказываний T^*_j определяется наилучшее высказывание по некоторому выбранному критерию y.

Этап 2. Рассматриваются все высказывания длины два типа $s=T^*_j/\backslash T_i$. Выбирается то высказывание $s^* = T^*_j/\backslash T^*_i$, для которого значение y максимально и т.д. Перебор заканчивается в том случае, если найдена закономерность или когда число элементарных высказываний, входящих в s*, превысило некоторое m_0. Если закономерность не найдена, критерий b, которые представляет собой процент экземпляров необучающей выборки, которые могут подойти логическому решающему правило увеличивается уменьшается и перебор повторяется

Этап 3. После нахождения первой закономерности s^w_1 исключаются объекты образа w, на которых выполнилось s^w_1 и процедура поиска закономерностей для оставшихся объектов продолжается до тех пор, пока не будет получено покрытие множества A^w.

Этап 4. Для остальных образов процедура поиска закономерностей аналогична [2].

После исключения признаков, вошедших в выбранные закономерности, можно на оставшихся признаках получить новый набор закономерностей и т.д. Критерий y, по которому оцениваются различные высказывания во время перебора, был выбран не случайно. Он подбирается таким образом, чтобы логическое решающее правило отбирало как можно больше своих высказываний и как можно меньше чужих. Интуитивно ясно, что чем больше для высказывания s величина Psw , которая представляет собой процент высказываний из образа w, которые подходят логическому решающему правилу, и меньше величина Prsw, представляющая собой процент высказываний не образе w, которые подошли логическому решающему правилу, тем предпочти-

тельнее его оставлять для дальнейшего перебора. Критерий должен быть таким, чтобы на начальных этапах перебора предпочтение отдавалось сочетанию, которое выполняется на большем числе реализаций образа w. Однако если ориентироваться в основном на величину Psw, то закономерность можно не получить.

Поэтому вес реализации образа w возрастает по мере увеличения номера этапа. С учетом изложенного в качестве критерия y была использована величина y=Psw - m/2* Prsw, где m – номер этапа перебора.

Данный алгоритм реализован на языке программирования Java. Выбор языка Java обусловлен общими нестрогими положениями: он наиболее адаптированный для программирования в различных операционных системах, в том числе и для программирования мобильных приложений.

Алгоритм моделирования кластеров

Пусть наш кластер состоит из n объектов, измеренных в m наименованиях, который могут принимать s значений признаков. Например, объект мальчик имеет признак «цвет глаз», который может принимать значение только «зеленый», «карий» и «голубой». Чтобы создать первый кластер, мы прибегаем к автозаполнению посредством распределения вероятности случайных величин, где каждый из s признаков в равной степени может быть выбран признаком объекта. Таким образом, если наш признак измерен в 4-х наименованиях, то вероятность выпадения каждого из наименований будет 25%.

Для моделирования второго кластера мы у i-го признака изменяем распределение вероятности. Например, пусть у i-го признака, который измерен в 3 наименованиях вероятность выпадения первого наименования будет 80%, второго – 10% и третьего, соответственно, 10%.

Для третьего и последующих кластеров алгоритм моделирования аналогичен.

Результаты обучения

В программе мы моделируем два кластера, задавая им параметры: количество элементов, количество характеристик и номера характеристик, которые будут отличаться по распределению вероятности от остальных параметров кластера.

После создания кластеров для первого кластера строится логическое решающее правило таким образом, чтобы логическое решающее правило подходило всем элементам обучающей выборки, но наименьшему количеству элементов второго кластера.

Чтобы проверить работоспособность программы необходимо проверить работу логического решающего правила на тестирующей выборки, в которой мы заранее знаем элементы и однозначно можем определить должно сработать логическое решающее правило на этом элементе или нет. На основе работы логического решающего правила на тестирующей выборке мы можем определить процент ошибки нашего правила.

Далее приведены таблицы зависимости ошибки от количество экземпляров обучающей выборки, количества параметров, вероятность распределения которых изменена и количества возможных значений в каждом из кластеров для количества параметров равное 20(табл.1), 50(табл.2) и 70 (табл.3).

Таблица 1 – количество ошибок в процентах от общего числа при 20 параметрах для каждого экземпляра обучающей выборки

Количество экземпляров в выборке	200		500		1000		2000		10000	
Количество возможных значений в каждом из столбцов	5	10	5	10	5	10	5	10	5	10
1	13,50%	15%	8,80%	13,60%	14,50%	7,70%	10,55%	16,20%	10,21%	7,38%
2	12,50%	16,50%	7,80%	11,80%	15,60%	7,70%	11,15%	16,10%	10,55%	7,52%
5	11,50%	16%	9%	10,80%	14%	8,20%	10,10%	16,80%	10,25%	7,23%

Таблица 2 – количество ошибок в процентах от общего числа при 50 параметрах для каждого экземпляра обучающей выборки

Количество экземпляров в выборке	200		500		1000		2000		10000	
Количество возможных значений в каждом из столбцов	5	10	5	10	5	10	5	10	5	10
1	13,40%	15%	8,85%	13,65%	14,75%	7,73%	10,85%	16,75%	10,19%	7,33%
2	12,60%	16,45%	7,70%	11,50%	15,46%	7,50%	11,25%	16,15%	10,57%	7,52%
5	11,45%	16%	9%	10,30%	15%	8,30%	10,10%	16,80%	10,21%	7,29%

Таблица 3 – количество ошибок в процентах от общего числа при 70 параметрах для каждого экземпляра обучающей выборки

Количество экземпляров в выборке	200		500		1000		2000		10000	
Количество возможных значений в каждом из столбцов	5	10	5	10	5	10	5	10	5	10
1	13,60%	16%	8,65%	13,25%	14,20%	7,90%	10,95%	16,55%	10,22%	7,34%
2	12,60%	16,40%	7,45%	11,65%	15,35%	7,65%	11,75%	16,45%	10,52%	7,57%
5	11,45%	16%	9%	10,35%	15%	8,60%	10,15%	16,25%	10,24%	7,35%

Как видно из таблиц, процент ошибки не сильно меняется при росте количества параметров для каждого экземпляра обучающей выборки. Объясняется это тем, что для построения логического правила однозначно определяющего принадлежность элемента к кластеру достаточно 10-15 параметров.

Также можно заключить, что чем больше элементов в обучающей выборке, тем меньше процент ошибок на тестирующей выборке. При возрастании количества возможных значений каждого из параметров также наблюдается уменьшения процента ошибок. Это происходит из-за того, что логическое решающее правило становится более уникальным и, соответственно, лучше классифицирует элементы тестирующей выборки.

Также при увеличении количества параметров с неравномерным распределением вероятности значений на «альтернативной» выборке наблюдается, как правило, уменьшение ошибки, так как логическое решающее правило становится более уникальным.

Литература

1. Антамошкин А.Н., Масич И.С. Выбор логических закономерностей для построения решающего правила распознавания. 2014. 5 с.
2. Лбов Г.С. Методы обработки разнотипных экспериментальных данных, 1981. 413 с.

Т.О. Куцаева

СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ ПО ЭМОЦИОНАЛЬНОЙ ОКРАСКЕ

(Самарский университет)

Задача классификации текстов - одна из самых важных в компьютерной лингвистике. Она позволяет решать задачи определения эмоциональной окраски высказывания, жанра текста и многие другие [1].

В ходе данной работы разработаны интеллектуальные системы распознавания эмоциональной окраски текста на основе логистической регрессии и многослойного персептрона с одним скрытым слоем. Логистическая регрессия обучается с помощью метода наименьших квадратов, для персептрона же использован алгоритм ADAM [2]. Обучение происходило в 5 эпох, размер каждой пачки — 32 примера.

Логистической регрессией называют статистическую модель, предназначенную для предсказания вероятности возникновения какого-то события с помощью подгонки данных к логистической кривой, представленной на рисунке 1 [3].

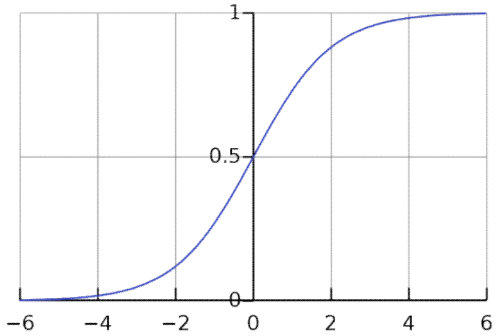

Рисунок 1 - Логистическая кривая, относительно которой делаются прогнозирования

Персептрон — это математическая или компьютерная модель восприятия информации мозгом. Появившись в конце 50-х - начале 60-х гг. XX века, он стал одним из самых первых моделей нейронных сетей [4]. На рисунке 2 представлена схема многослойного персептрона с одним скрытым слоем.

В данной работе в качестве набора данных используется Twitter — набор из 1,6 млн случайно выбранных уникальных сообщений социальной сети