УДК 004.832

Т.В. ГАВРИШ, С.Н. ЖИВОЛУН

АВТОМАТИЗИРОВАННАЯ ОЦЕНКА КАЧЕСТВА ПОДГОТОВКИ ВЫПУСКНИКОВ ВУЗА ПОСРЕДСТВОМ МЕТОДА МНОГОМЕРНОЙ КЛАССИФИКАЦИИ

Предлагается использование интегрированных показателей качества, формируемых в пространстве сведений, содержащихся в отзывах-характеристиках. Выполняется реинжиниринг задачи автоматизированной оценки качества подготовки выпускников вуза. Задача решается посредством метода автоматической классификации (кластер-анализа) и байесовских процедур.

1. Проблема повышения качества подготовки выпускников вуза в современных условиях

Очевидно, что в настоящее время эффективность экономики, научно-технический уровень производства и социально-экономический прогресс в целом в значительной мере зависят от состояния системы образования и объема накопленных обществом знаний. Это обстоятельство определяет необходимость подготовки грамотных специалистов, квалификация которых соответствует современным достижениям НТП.

Коммерциализация экономики вносит существенные изменения в систему трудоустройства выпускников, акцентируя внимание на качестве обучения, соответствии уровня профессиональной подготовки требованиям потенциальных мест работы, а также на успешном противодействии вузам-конкурентам. Положительное решение этих вопросов утверждает за выпускниками вуза репутацию грамотных, квалифицированных и, значит, привлекательных для работодателей субъектов рынка труда. Это обстоятельство весьма существенно при формировании госзаказа вузу и наборе контингента студентов, обучающихся по контрактной форме.

Для постоянного отслеживания потребностей целевого рынка в специалистах, обладающих необходимым объемом и уровнем знаний, в ХНУРЭ организована стажировка выпускников. Её суть состоит в рассылке по местам работы выпускников отзывов-характеристик (ОХ) с набором вопросов, оценивающих показатели качества подготовки и работы стажеров. По результатам стажировки формируются управляющие воздействия, адаптирующие структуру учебного процесса и содержание отдельных дисциплин к запросам предприятий-потребителей выпускников университета.

2. Задача оценки качества подготовки выпускников вуза и основные принципы ее решения

Данная задача была автоматизирована в рамках информационной аналитической системы «Университет» (модуль «Кафедра»). Следует отметить, что в настоящее время результатом статистической обработки поступивших ОХ являются сведения о количестве выпускников, которых можно характеризовать отдельными показателями качества подготовки и работы на предприятии (качество выполнения поручений, общетеоретический уровень подготовки, уровень конкретных знаний и т.д.). Вместе с тем известно [1], что выводы, полученные в результате анализа информации, поступившей от множества статистически обследованных объектов (в нашем случае выпускников), должны опираться одновременно на совокупность всех показателей с обязательным учетом структуры и характера их взаимосвязей. Данное заключение позволяет считать, что для объективной и полной оценки качества подготовки выпускников необходим комплексный учет всех показателей, содержащихся в ОХ, и статистических связей между ними.

Выработке подобных интегрированных показателей и технологии их применения при статистической обработке поступивших ОХ посвящена настоящая статья.

В указанной постановке решаемая задача может быть сведена к построению типологии, т.е. выделению из исходной совокупности объектов–выпускников определенных типов (групп), однородных в содержательном смысле и в некотором отношении похожих друг на друга. Известно [2], что для построения типологии могут быть использованы методы многомерной классификации, в соответствии с которыми определяется, к какому из типов принадлежит каждый из объектов исходной совокупности. В случае удачно выбранного алгоритма классификации полученные классы (кластеры) полностью соответствуют изначально заданным типам. Последнее предполагает наличие априорно заданных моделей типов классифицируемых объектов.

С учетом изложенного выше содержательная трактовка задачи состоит в разработке процедур разделения контингента выпускников в зависимости от качества их подготовки на несколько типов. Основанием для отнесения того или иного выпускника к одному из типов являются результаты статистической обработки поступивших ОХ, в процессе которой следует использовать только наиболее информативные показатели. Данное замечание определяет также необходимость снижения размерности информационного пространства, образованного совокупностью содержащихся в ОХ сведений. Отметим [1], что в подавляющем большинстве случаев процедуры классификации и снижения размерности выполняются в рамках одной задачи.

Тогда результатом решения задачи является описание типов в пространстве наиболее информативных показателей качества подготовки и установление правил отнесения объектов-выпускников к одному из них. Следует отметить, что выбор информативных показателей и их градаций фактически задает перечень и описание структурных единиц входных сообщений, характеристика которых необходима для разработки модели базы данных задачи.

Во многих случаях довольно трудно идентифицировать объекты по одному или даже нескольким показателям (признакам) и выполнить их однозначную классификацию. Это связано с социально-экономической природой объектов классификации и конкретно со следующими обстоятельствами. Во-первых, в ОХ при оценке подготовки выпускников превалируют качественные признаки. Во-вторых, границы между градациями качественных признаков строго установить невозможно, что приводит к их определению на основе субъективных восприятий (например, основательный, средний, низкий уровни теоретической подготовки). И, наконец, в-третьих, отсутствие строгой формализации цели классификации и замена ее содержательной интерпретацией приводит к неоднозначности результатов решения задачи.

3. Применение метода многомерной классификации для решения поставленной задачи

Учитывая изложенное выше, задачу можно формализовать следующим образом. В результате статистической обработки отзывов-характеристик n объектов-выпускников $B_1, B_2, ..., B_n$ получена исходная информация в виде матрицы «объект-свойство»:

$$X = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(p)} & x_2^{(p)} & \dots & x_n^{(p)} \end{bmatrix}, \qquad (1)$$

где $x_i^{(l)}$ – значение l-го признака, $l = \overline{1,p}$, в ОХ на i-го выпускника, а i-й столбец матрицы $X_i = (x_i^{(1)}, x_i^{(2)}, ..., x_i^{(p)})$ характеризует качество подготовки выпускника B_i, т.е. представляет результат заполнения предприятием ОХ на этого стажера.

Тогда задача классификации заключается в том, чтобы всю анализируемую совокупность выпускников $B=\{B_i\}$, $i=\overline{1,n}$, статистически представленную в виде матрицы (1), разбить на сравнительно небольшое число однородных, в определенном смысле, классов. Для формализации задачи разумно интерпретировать отдельные объекты в виде точек в соответствующем признаковом пространстве. В рассматриваемом случае эти точки являются непосредственным геометрическим изображением наблюдений X_i, $i=\overline{1,n}$ в p-мерном пространстве. Сгустки точек образуют классы (кластеры), а объекты-выпускники, принадлежащие одному кластеру, должны обладать некоторой мерой сходства, которая трактуется как сравнительно одинаковый уровень подготовки.

Из этого вытекает, что при формировании классов следует придерживаться принципа обеспечения наибольшего сходства объектов внутри классов и возможно большего различия между ними.

При использовании метода кластерного анализа понятие сходства объектов, фактически определяющее их однородность, формализовано [1]. Геометрическая близость точек признакового пространства характеризуется расстоянием $d(B_i,B_j)$ между объектами B_i и B_j из исследуемой совокупности или степенью сходства $r(B_i,B_j)$ этих же объектов, что соответствует агломеративным кластер-процедурам.

В первом, наиболее распространенном, случае метрика признакового пространства задается с помощью евклидового расстояния:

$$d_{ij}^2=\sum_{l=1}^{p}\left|x_i^{(l)}-x_j^{(l)}\right|^2,\ i,j=\overline{1,n};\ l=\overline{1,p}. \qquad (2)$$

Однако если учесть качественный характер признаков, описывающих множество B_i, то для использования (2) необходимо провести ранжирование каждого из них. Последнее с учетом субъективности данной процедуры ставит под сомнение достоверность классификации, поскольку однородность кластеров непосредственно зависит от ранга каждого из $x_i^{(l)}$. Действительно, невозможно объективно оценить «вес» каждого из показателей качества подготовки выпускников.

Кроме того, довольно часто (в том числе и в рассматриваемой задаче) принадлежность объектов к определенным типам обуславливается не близостью в среднем по ряду признаков, а их совпадением по некоторым ключевым признакам. При этом объекты, различающиеся хотя бы по одному ключевому признаку, могут принадлежать различным типам. В таких условиях традиционные кластер-процедуры не продуктивны и в [2,3] рекомендуется использовать специальные методы группировок. Впрочем, последние, согласно [3], могут быть отнесены к иерархическим кластер-процедурам. Суть этих методов состоит в использовании формальных критериев значимости различных комбинаций признаков и автоматизации их поиска на основе определенных алгоритмов. Данный подход, а именно, метод последовательных разбиений, принят при оценке качества подготовки выпускников.

Критерием значимости признака выбрана его информативность. Действительно (и это также согласуется с интуитивными представлениями), наиболее существенным является тот признак, значение которого позволяет с большей уверенностью, по сравнению с другими признаками, судить как об объекте, так и об остальных признаках. Тогда информативность признака $x^{(l)}$, $l=\overline{1,p}$ может быть определена мерой Шеннона, а его суммарная взаимосвязь со всеми остальными признаками оценена изменением энтропии объекта B_i, $i=\overline{1,n}$ после его классификации по этому признаку. Средневзвешенное значение

показателя взаимосвязи признака $x^{(l)}$ со всеми остальными признаками $x^{(q)}$ равно, согласно [2],

$$M_l = (\sum_{q \neq l} h_{q/l}) / H_{B_i}, \qquad (3)$$

где $h_{q/l}$ – условная энтропия объекта по признаку q после выяснения его состояния по признаку l; H_{B_i} – исходная энтропия объекта B_i.

Если рассчитать для каждого признака значение M_l, то направление оптимального разбиения на первом шаге будет определяться $\max_{(l)} M_l$, т.е. следует выбрать наиболее информативный признак. Если таких признаков несколько, то выбор производится по содержательным соображениям.

Таким образом, можно заключить, что построение типологии множества B осуществляется последовательно, согласно упорядоченным по (3) признакам. Вначале образуются классы в соответствии с градациями наиболее информативного признака. Затем каждый из полученных классов разбивается на подклассы градациями следующего по уровню информативности признака и т.д. Для практического использования алгоритма классификации необходимо указать правило останова процесса. Согласно [2,3], им может быть условие образования классов из полностью однородных элементов, т.е. дальнейшее разбиение нецелесообразно при достижении заданной степени однородности составляющих подмножеств объектов.

Для оценки однородности формируемых классов может быть использована вероятность ошибочной классификации. Обозначим через $r_l(k)$ величину, равную модальной частоте проявления признака $x^{(l)}$ в k-м классе, т.е. равную наибольшему числу объектов, характеризующихся одной из градаций этого признака, среди всех n_k объектов k-го класса. Тогда оценкой вероятности отнесения B_i по признаку $x^{(l)}$ к градации, отличной от модальной в классе, будет величина $P_{(l)}(\varepsilon) = (n_k - \sum_k r_l(k)) / n_k$, осреднив которую по всему множеству признаков, получим оценку средней вероятности ошибочной классификации:

$$P(\varepsilon) = (\sum_l P_l(\varepsilon)) / p. \qquad (4)$$

Правило останова процесса образования классов может быть построено на основе сравнения (4) с некоторой пороговой величиной $P_П(\varepsilon)$. При $P(\varepsilon) \leq P_П(\varepsilon)$ можно считать достигнутым заданный уровень однородности классификации в целом. В рассматриваемой задаче было принято $P_П(\varepsilon) = 0{,}2$.

Наименование полученных типов – это последовательность наименований градаций признаков, упорядоченная согласно последовательности шагов разбиения. Например, в класс $[x_1^{\,2}; x_2^{\,2.3}; x_4^{\,1}]$ входят объекты, имеющие вторую градацию признака 1, вторую и третью градации признака 2 и первую градацию признака 4. Признак 3 оказался неинформативным и в классификации не участвовал. Отметим, что в процессе построения типологии происходит также сокращение размерности признакового пространства. Однако это не означает исключение не вошедших в наименование типов признаков из исходного

статистического материала (выборки ОХ). Последние могут быть полезны при построении типологий в иных аспектах.

Построенная типология позволяет идентифицировать объекты B_i с учетом содержащейся в ОХ информации, т.е. определить степень принадлежности (подмножества размыты) каждого объекта к выделенным типам. Предположим, что совокупность объектов B состоит из $S_k, k=\overline{1,m}$ классов, полученных согласно работе алгоритма последовательных разбиений. Тогда каждому из объектов поставим в соответствие m действительных неотрицательных чисел $P(S_k/B_i)$, означающих вероятность отнесения объекта B_i с признаками, описываемыми согласно (1), к классу S_k. Естественно потребовать, чтобы $\sum_{k=1}^{m} P(S_k/B_i)=1$. Тогда оценка $P(S_k/B_i)$ может быть определена по формуле Байеса следующим образом:

$$P(S_k/B_i)=[P(S_k)P_k(B_i)]/[\sum_{k=1}^{m} P(S_k)P_k(B_i)], \qquad (5)$$

где $P(S_k)=n_k/n$ – вероятность «заполненности» каждого k-го класса; $P(B_i)=\prod_{l} P_k(x_i^l)$ – вероятность принадлежности объекта к k-му классу с учетом всей совокупности признаков из описания этого класса. Методика расчета указанных вероятностей приведена в [3].

Выражение (5) является решающим правилом, пользуясь которым можно определить степень принадлежности объекта к типу с известным наименованием.

4. Выводы

В рассматриваемой задаче было выделено три типа качества подготовки выпускников: достаточный, приемлемый, слабый. В результате работы алгоритма последовательных разбиений пространство показателей качества подготовки выпускников, предусмотренное ОХ, сокращается до трех признаков со следующими градациями:

– уровень конкретных знаний по специальности (достаточный, средний, недостаточный);

– наличие навыков практической деятельности (достаточные, недостаточные, отсутствуют);

– наличие навыков НИР (достаточные, недостаточные, отсутствуют).

Наименования указанных типов приведены в таблице.

Тип качества подготовки	Признак		
	Конкретные знания	Практические навыки	Навыки НИР
ДП	Д или С	Д или Н	Д
ПП	Д или С	Д или Н	Н
СП	Н	Н или О	О

В ней приняты следующие обозначения: ДП - достаточная подготовка; ПП - приемлемая подготовка; СП - слабая подготовка; Д - достаточные; С - средние; Н - недостаточные; О - отсутствуют.

Идентификация выпускников 2001 г. специальности ИУСТ, согласно (5), по результатам статистической обработки 24 ОХ дала следующую классификацию: 12 выпускников имеют достаточный уровень профессиональной подготовки; 8 – приемлемый; 4 – слабый.

Следует отметить, что сформированное признаковое пространство непосредственно связано с логическим моделированием данных в задаче автоматизированной оценки качества подготовки выпускников. В частности, каждая запись базы данных соответствует объекту-выпускнику, поля должны содержать все фикси-

руемые в ОХ показатели-признаки, а тип шкалы, используемой для их оценки, - возможные значения соответствующих структурных единиц входных сообщений.

Список литературы: 1. *Прикладная* статистика: Классификация и снижение размерности / Под ред. *С.А.Айвазяна*. М.: Финансы и статистика, 1989. 607с. 2. *Типология* и классификация в социологических исследованиях / Под ред. *В.Г.Андреенкова*. М.: Наука, 1982. 296 с. 3. *Елисеева И.И., Рукавишников В.О.* Группировка, корреляция, распознавание образов. М.: Статистика, 1977. 144 с.

Поступила в редколлегию 25.01.2003

Гавриш Татьяна Валентиновна, канд. техн. наук, доцент кафедры ИУС ХНУРЭ. Научные интересы: методы статистической обработки многомерной информации. Адрес: Украина, 61023, Харьков, ул. Мироносицкая, 99, кв. 30. тел. 43-69-33.

Живолун Сергей Николаевич, студент группы ИУСТ-99-2 кафедры ИУС ХНУРЭ. Научные интересы: проектирование информационных систем. Адрес: Украина, 61047, Харьков, ул. Электровозная, 4, кв.99.