

ИНФОРМАЦИОННЫЙ ПОДХОД ВЫДЕЛЕНИЯ СУЩЕСТВЕННЫХ ФАКТОРОВ В АНАЛИЗЕ ЭМПИРИЧЕСКОЙ ИНФОРМАЦИИ

В. Г. Вартанова

Проблема совершенствования технологического процесса исследования, автоматизации его на основе средств вычислительной техники в смысле автоматизации сбора, обработки, интерпретации и анализа эмпирической базы исследований, формирование эффективных справочных систем по результатам исследований были и остаются актуальными и по сей день.

После получения эмпирической базы исследований возникает задача дать общую характеристику объекта исследования и его структуры, выделить «естественные» компоненты, установить, какие соотношения между ними являются существенными с точки зрения поставленных целей, а какие нет, описать направленность развития объекта и на этой основе спрогнозировать его будущую структуру.

Поскольку в реальности объекты социально-экономического исследования представляют собой исключительно сложные, многокомпонентные системы, то для них трудно, а порой и практически невозможно учесть все возможные виды взаимодействий и рассчитать возможные исходы тех или иных событий. Поэтому одной из актуальных проблем в социально-экономических исследованиях является проблема выявления, исследования и интерпретации зависимостей всех признаков, которые характеризуют различные социально-экономические явления. Данная проблема возникает на всех этапах исследования, начиная с формирования выборочной совокупности и заканчивая разработкой рекомендаций по управлению исследуемыми процессами, т. е. заканчивая конечной целью проведения любого исследования. Причем, эта проблема слишком сложна в силу объема измеряемой информации. Поэтому первой из задач, возникающих при обработке результатов экспериментальных данных, является задача выделения существенных, наиболее информативных факторов, что позволит сократить описание измеряемой информации.

Одним из перспективных подходов к решению сформулированных выше задач является подход, основанный на использовании корреляционной (ковариационной) матрицы. Данный подход опирается на представление об экстремальной группировке параметров [1; 2]. Но его трудно реализовать, поскольку возникают трудности при работе с параметрами, которые измерены в различных шкалах, а также трудности из-за ограниченного объема выборки.

При решении данного рода задач интерес представляет информационный подход при выделении существенных факторов, который является развитием работ [1; 2; 3] по реализации метода множественных сравнений непересекающихся групп объектов с проблемной группой для анализа связи.

Как отмечено в [1], разбиение факторов на группы проводится по сильной в некотором смысле корреляции между факторами в каждой из групп и относительно слабой между факторами разных групп. При этом предполагается, что такие взаимосвязанные группы факторов косвенно измеряют интересующие свойства изучаемого объекта, не поддающегося прямому измерению.

Применение теоретико-информационного подхода позволяет расширить процедуру исследования, решить задачу выделения групп факторов, которые в наибольшей степени характеризуются внутренней взаимосвязанностью и взаимообусловленностью и определить структурную упорядоченность и детерминацию исследуемой системы.

Рассмотрим информационную меру, позволяющую сравнивать между собой различные группы факторов с точки зрения их упорядоченности и введем критерий выбора наиболее информативных групп.

Пусть исследуемая система характеризуется эмпирическим множеством конечного объема (экспериментальными данными). Каждый объект множества определяется m -мерным вектором результатов измерения каждого признака. Из множества признаков выделяется то или иное подмножество (группа) признаков (например, группа по уровню дохода, группа по жилплощади, группа по покупательскому спросу на тот или иной вид товара и другие группы, которые могут быть как пересекающимися, например, уровень дохода и жилплощадь, так и не пересекающимися). Для этого выделенного подмножества определяется величина энтропии как для каждого признака в отдельности, так и совместная энтропия всей группы признаков [3]. Согласно теореме Хинчина [3], энтропия системы в целом уменьшается вследствие взаимодействия ее подсистем. Уменьшение же энтропии эквивалентно росту негэнтропии, т. е. накоплению информации и увеличению степени упорядоченности системы. Поэтому, увеличивая размер признакового пространства до тех пор, пока интенсивность роста негэнтропии при увеличении длины набора признаков преобладает над интенсивностью роста энтропии подсистемы признаков, получаем наиболее информативную систему признаков с наибольшей структурной упорядоченностью и внутренней взаимосвязанностью ее элементов. Полученные таким образом подмножества признаков являются наиболее информативными внутри каждой из групп параметров.

Данный информационный подход к решению задачи выделения существенных факторов на основе экстремальной группировки параметров дает возможность построить экстремальные вычислительные процедуры для обработки разнотипных признаков в условиях высокой неопределенности.

Литература

1. Дзмида Г. К вопросу об экстремальной группировке параметров. – В сб. «Теория оптимальных решений», вып.8, Вильнюс, 1992. С. 46–53.
2. Ростовцев П. С., Костин В. С., Олех А. Л. Множественные сравнения в детерминанном и типологическом анализе. //Анализ и моделирование экономических процессов переходного периода в России. Выпуск 3.-Новосибирск, ИЭиОПП СО РАН, 1999. С. 148–164.
3. Хинчин А. Я. Понятие энтропии в теории вероятностей. – «Успехи математических наук», 1992, т.ХІІІ, вып. 3(55). С. 3–20.

СИНТАКСИЧЕСКИЙ АНАЛИЗ КАК СПОСОБ СНЯТИЯ ОМОНИМИИ ПРИ АВТОМАТИЧЕСКОЙ ОБРАБОТКЕ ТЕКСТА

Ю. А. Володькина

Разрешение омонимии является одной из самых главных задач исследователей и разработчиков интеллектуальных систем, связанных с естественным языком.

Существующие подходы к снятию омонимии разделяются на вероятностные, использующие статистику встречаемости грамматических признаков слов в больших корпусах, и надетерминированные, т. е. основанные на синтаксическом анализе и синтаксических словарях. Синтаксическому анализу предшествует морфологический анализ. Он представляет собой обязательный начальный этап автоматического анализа для любого языка.

Автоматический морфологический анализ – это обработка отдельно взятых словоформ, в результате которой каждой словоформе ставится в соответствие морфологическая информация – характеристика, изображающая в явной и удобной форме те свойства словоформы, которые необходимы для последующего синтаксического анализа. К началу последнего анализируемый текст оказывается представленным в виде последовательности информации к словоформам, так что алгоритм синтаксического анализа имеет дело не со словоформами, а только с соответствующими информацией.

Синтаксический анализ – главная часть всей системы анализа. Это совокупность операций, которые выполняются над последовательностями информации, представляющими исходный текст, для установления синтакси-