студентам ниже, вследствие чего заочное обучение наиболее благоприятным представляется для получения второго и последующих образований.

Важнейшим направлением совершенствования современно системы образования может стать создание качественно новой системы дистанционного образования на основе быстро развивающихся информационных технологий.

Информационные технологии в образовании - это ответ на вызовы новой экономики, изменяющий роль обучения в современном мире и позволяющий сделать его непрерывным, массовым и эффективным. Дистанционное образование организуется так, что обучающийся имеет возможность пользоваться образовательными ресурсами и общаться с преподавателем в любое удобное для себя время в необходимом объеме. Место пребывания обучающегося не имеет значения, главное условие - наличие компьютера и доступа в Интернет. Соответственно, снимаются и возрастные ограничения. Дистанционное обучение дешевле очного в 2 - 4 раза, существующие средства связи позволяют учиться в любом учебном заведении планеты.

Развитый мир давно оценил указанные преимущества дистанционного образования. Сейчас примерно 1/5 всех образовательных услуг предоставляется через Интернет. Дистанционное образование присутствует почти в каждом серьезном западном вузе, даже если не является приоритетным.

Таким образом, современные белорусские предприятия уже осознали важность инвестиций в интеллектуальный капитал, выделяют средства для подготовки, переподготовки и повышения квалификации персонала. В то же время необходимо создание такой системы образования, которая сможет удовлетворить потребности современных организаций в высококвалифицированных работниках, способных стать ее интеллектуальным капиталом.

УДК 004.823

ВЫЯВЛЕНИЕ НЕЯВНЫХ ЗАКОНОМЕРНОСТЕЙ ПРИ ПОСТРОЕНИИ МОДЕЛЕЙ С ИСПОЛЬЗОВАНИЕМ СИСТЕМЫ SPSS CLEMENTINE

В.Л. Шарстнёв, Е.Е. Витрук, В.С. Соловьёва
УО «Витебский государственный технологический университет», г. Витебск, Республика Беларусь

В наше время во многих областях деятельности людям приходится сталкиваться с огромными массивами информации. В таких случаях провести какой-либо анализ очень сложно и порой невозможно в силу объёмов и неоднородности этой информации. Традиционные статистические методы подчас оказываются бессильны. Одним из перспективных направлений в анализе таких больших объёмов данных является применение технологии Data Mining.

Data Mining переводится как «добыча» или «раскопка данных». Согласно определению одного из основателей данного направления – Григория Пиатецкого-Шапиро, Data Mining — это процесс обнаружения в сырых данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности.

В основу технологии Data Mining положена концепция шаблонов, представляющих собой закономерности. В результате обнаружения этих скрытых от невооруженного глаза закономерностей решаются задачи Data Mining. Большинство авторитетных источников

относят к этим задачам классификацию, кластеризацию, прогнозирование, ассоциацию и последовательность.

Классификация – наиболее простая и распространенная задача Data Mining. В результате решения задачи классификации обнаруживаются признаки, которые характеризуют группы объектов исследуемого набора данных.

Кластеризация является логическим продолжением идеи классификации. Это задача более сложная, особенность кластеризации заключается в том, что классы объектов изначально не предопределены. Результатом кластеризации является разбиение объектов на группы.

Суть ассоциации в том, что ходе решения задачи поиска ассоциативных правил отыскиваются закономерности между связанными событиями в наборе данных.

Последовательность позволяет найти временные закономерности между связанными во времени событиями. Задача последовательности подобна ассоциации, но ее целью является установление закономерностей не между одновременно наступающими событиями, а между событиями, происходящими с некоторым определенным интервалом во времени.

В результате решения задачи прогнозирования на основе особенностей исторических данных оцениваются пропущенные или же будущие значения целевых численных показателей. Если удаётся найти шаблоны, адекватно отражающие динамику поведения целевых показателей, есть вероятность, что с их помощью можно предсказать и поведение системы в будущем.

Data Mining является мультидисциплинарной областью, возникшей и развивающейся на базе достижений прикладной статистики, распознавания образов, методов искусственного интеллекта, теории баз данных и других. Поэтому существует обилие методов и алгоритмов, реализованных в различных действующих системах Data Mining. К этим методам стоит отнести технический анализ, нейронные сети, системы рассуждений на основе аналогичных случаев, деревья решений, алгоритмы ограниченного перебора и некоторые другие.

Сфера применения Data Mining ничем не ограничена — она везде, где имеются какие-либо данные. Но в первую очередь методы Data Mining сегодня заинтересовали коммерческие предприятия, развертывающие проекты на основе информационных хранилищ данных. Опыт многих таких предприятий показывает, что отдача от использования Data Mining может достигать 1000%. Основные сферы применения технологии Data Mining – это розничная торговля, банковское дело, медицина, страховое дело, молекулярная и генная инженерия и другие. Особенность этих областей заключается в их сложной системной организации. Данные в указанных областях неоднородны, разнообразны, нестационарны и часто отличаются высокой размерностью.

Одним из наиболее известных Data Mining продуктов является Clementine компании SPSS (с последней версии название пакета изменено на PASW Modeler).

Clementine – это автоматизированное рабочее место для Data Mining, позволяющее быстро разрабатывать прогностические модели с привлечением бизнес-экспертизы, а затем внедрять полученные модели для усовершенствования процесса принятия решений. К преимуществам пакета относятся разработка программного комплекса на основе методологический подхода CRISP (ясный, четкий). являющегося стандартом де-факто в области Data Mining; наличие огромного числа специальных статистических методов, возможность предварительной подготовки данных, наличие четкой документации по работе с программой, отсутствие необходимости знания цели в начале исследования, визуальный подход к поиску решения.

Clementine используется для решения задач в следующих областях:

1. Общественная сфера. Правительства используют Clementine для исследования огромного количества данных, выявления случаев мошенничества, таких как отмывание денег и уклонение от уплаты налогов

2. CRM (Customer Relationship Management) – система управления взаимодействия с клиентами. Улучшения могут быть достигнуты, благодаря правильной классификации типов клиентов.

3. Web-Mining. Мощные алгоритмы Clementine позволяют точно определить, что именно люди делают на web-сайтах и доставлять им именно ту информацию или продукты, которые им необходимы.

4. Фармацевтика и биоинформатика. В области биоинформатики программа применяется для определения генетических факторов, влияющих на заболевания.

Для более наглядного представления о работе программного пакета SPSS Clementine следует рассмотреть пример задачи, решаемой с применением алгоритма списка решений – одного из методов Data Mining. Данный алгоритм генерирует правила, показывающие большую или меньшую вероятность дачи положительного или отрицательного ответа. Пример основан на анализе деятельности компании, которая хочет достичь более прибыльных результатов в будущих маркетинговых компаниях путём соотнесения подходящего маркетингового предложения с каждым клиентом. Модель, исходя из данных о прошлых кампаниях, определяет характеристики клиентов, которые с наибольшей степенью вероятности примут те или иные предложения. На основании результатов составляется список этих клиентов, и потом им рассылается необходимая информация относительно новой кампании.

Исходный файл содержит информацию о предложениях клиентам, характеристики клиентов и их реакцию на предложения со стороны компании. Целевым показателем будет являться реакция клиента (response), т.е. примет он или не примет предложение. Количество характеристик клиентов ограничено лишь информацией, имеющейся в распоряжении у компании.

Весь процесс отыскания данных будет выполняться с помощью так называемых узлов – логических элементов программы, выполняющих определённые действия и процессы и следующих один за другим. Узлы соединяются стрелками, что определяет последовательность выполнения программой операций. Все соединённые между собой узлы представляют собой поток решения (рисунок 1), на входе которого – исходная информация, а на выходе – найденные скрытые закономерности.

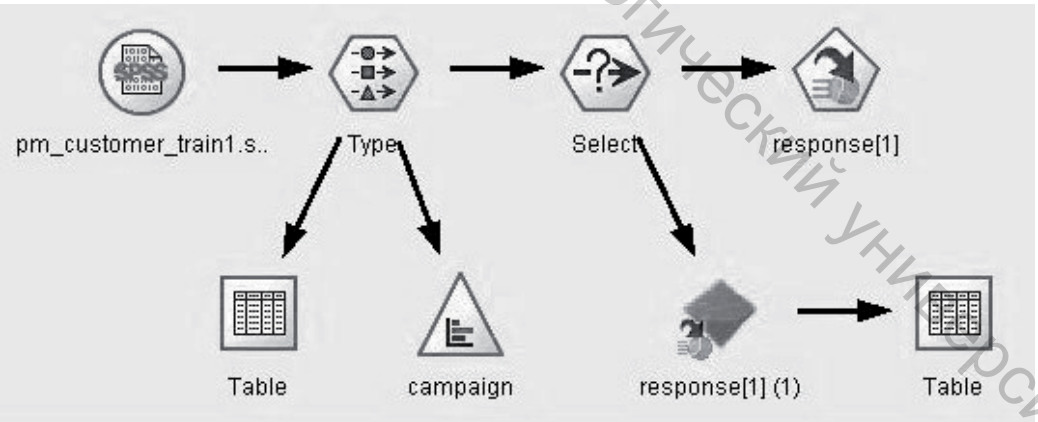

Рисунок 1 - Поток решения задачи

Для построения модели первоначально необходимо добавить в поток файл данных, содержащий всю информацию о клиентах. Для того, чтобы определить типы и характеристики всех переменных, находящихся в исходном файле, необходимо соединить исходный файл с узлом Type. Чтобы вывести всю информацию о клиентах в табличной форме, необходимо к узлу Type добавить выходной узел Table.

Исходный файл содержит информацию о нескольких рекламных кампаниях, однако для получения более точного результата необходимо анализировать результаты одной из

кампаний. Для того, чтобы определить, какая из кампаний была наиболее масштабной, т.е. была направлена на наибольшее количество клиентов, к узлу Type необходимо добавить узел Distribution (в примере имя узла – «Campaign»), который позволит определить кампанию, затронувшую максимальное количество клиентов. Наибольший удельный вес занимает маркетинговая кампания №2. Поэтому из всего массива данных должны быть отобраны только те записи, где параметр «Campaign» равен 2 (узел Select). Далее будет осуществляться анализ только этих записей.

После этого добавляем к потоку узел Decision List (в примере имя узла – «Response»), который и является основным инструментом в данном примере. Для начала целесообразно просто подсчитать, сколько записей будет подвергаться анализу и в какой части из них значение целевого показателя будет равно 1. Также мы определяем количество сегментов, на которые будут разбиваться все анализируемые записи о клиентах. Для простоты примера примем этот показатель равным трём. После этого нужно запустить анализ, выбрав Launch Interactive Session. После этого появляется окно с данными о том, сколько человек участвовало в кампании под номером 2 и какой процент из них принимал предложение фирмы. В нашем случае из 13504 человек предложение приняли 1952, что составляет 14,45%.

После всех этих действий с помощью команды Start Default Mining Task мы запускаем процесс поиска скрытых закономерностей среди отобранных записей. В итоге система выдаёт нам три альтернативных модели (это указанное нами ранее количество сегментов) характеристик клиентов, которые с наибольшей вероятностью принимали маркетинговое предложение №2. Эти показатели представляются в виде неравенств (например, «доход>55267»). Стоит отметить, что вероятность принятия предложения среди клиентов данных сегментов выше средней и может превышать 90% против общих 15%. Данная модель охватывает лишь небольшую часть клиентов, однако уже это позволяет значительно повысить экономическую эффективность маркетинговых кампаний.

Далее с помощью команды Organize mining tasks/Down Search среди оставшихся записей можно определить правила, характеризующие клиентов, которые с наименьшей вероятностью принимали предложения фирмs. Кроме этого, записи, в которых вероятность равна нулю исключаются из анализа.

Чтобы получить более яркое представление о том, насколько полезна получившаяся модель, можно измерить определённые экономические показатели для клиентов определенных сегментов. Интеграция с MS Excel позволяет рассчитать такие показатели, как прибыль и предельная прибыль для каждого из оставшихся сегментов.

Продолжая далее производить анализ модели пользователь может экспериментировать с различными вариантами поиска закономерностей, редактировать и создавать свои сегменты и изменять модель различными способами. Когда пользователь удовлетворен результатами, он может сгенерировать искомую модель для того, чтобы использовать её далее в своих целях для достижения большего экономического эффекта.

Таким образом, нетрудно увидеть, как применение технологии Data Mining, в частности с применением пакета SPSS Clementine, позволяет находить скрытые закономерности в больших объёмах данных. Данная технология очень перспективна, потому что она позволяет при относительно малых затратах физического и интеллектуального труда, осуществлять ту работу, которая без её применения была бы невозможна.