21. ВОПРОСЫ И ПРОБЛЕМЫ ПОДГОТОВКИ, ОБРАБОТКИ И ПРЕДСТАВЛЕНИЯ ДАННЫХ ДЛЯ КОМПЬЮТЕРНОГО И ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА

Юркевич Н.В.

Белорусский государственный университет информатики и радиоэлектроники г. Минск, Республика Беларусь

Шинкевич Е.А. – канд. физ.-мат. наук

Аннотация. В работе освещаются такие аспекты, как понятие интеллектуального анализа данных, селективные смещения и предвзятость в обработке информации, статистические парадоксы и феномены, а именно парадокс Симпсона и феномен Уилла Роджерса. Также обсуждается явление джеррименденинга и его влияние на результаты анализа данных. Особое внимание уделяется манипуляциям визуализацией данных и их воздействию на восприятие и интерпретацию информации.

Интеллектуальный анализ данных – это компьютеризованная технология, используемая в аналитике для обработки и исследования больших наборов данных. В основе интеллектуального анализа данных, как и в Data Mining, лежит идея активного применения математических методов, таких как оптимизация, генетические алгоритмы, распознавание образов, статистика, и т.д., а также использующих визуальное представление информации. Благодаря интеллектуальному анализу данных можно достичь следующих результатов: описать имеющиеся данные или сделать прогнозы на будущее. Используя инструменты и методы интеллектуального анализа данных, организации могут выявлять шаблоны и отношения, скрытые в данных. Интеллектуальный анализ данных преобразует необработанные данные в практические знания. Компании используют знания для решения проблем, анализа будущего влияния бизнес-решений и повышения прибыли [1].

Однако уже на этапах сбора данных для анализа можно встретить различные проблемы, так как для дальнейшего успеха необходимо собрать актуальные, качественные и релевантные наборы данных, получить права доступа, либо в случае отсутствия подобных ресурсов, самостоятельно провести исследования, опросы и анкетирования. Рассмотрим, какие именно трудности могут возникнуть на данном этапе.

Селективные смещения и когнитивная предвзятость

- Ошибка отбора - выводы, сделанные применительно к какой-либо группе, могут оказаться неточными вследствие неправильного отбора в эту группу.

Ни один добровольный опрос не может предоставить полную репрезентацию каждого члена в обществе.

- Специальный отбор – оставить только прецеденты, подтверждающие теорию.
- Самоселективные смещения: когда объекты выборки оценивают/выбирают себя же.
- Когнитивные смещения связаны с особенностями человеческого восприятия, например ошибка воспоминания.
- Предвзятость наблюдателя: склонность к подтверждению своей точки зрения.
- Предвзятость ответа. В том числе эффект социальной желательности: опрашиваемые дают ответ, за который их не осудят. Это причина анонимных опросов и неявного сбора информации [2].

Предотвратить полностью возникновение подобных смещений и ошибок невозможно, однако мы можем их снизить, правильно составляя опросы и отбирая данные, еще не подверженные таким когнитивным смещениям как ошибка воспоминания.

Следующим этапом в нашем анализе станет обработка данных и интерпретация результатов. В данном случае нам стоить помнить о существовании подобных **статистических парадоксов и феноменов**:

Парадокс Симпсона (также Парадокс Юла—Симпсона или «парадокс объединения») — явление в статистике, когда при наличии двух групп данных, в каждой из которых наблюдается одинаково направленная зависимость, при объединении этих групп направление зависимости меняется на противоположное. Одним из наиболее известных примеров парадокса Симпсона является случай половой дискриминации при поступлении в Калифорнийский университет Berkeley. При подсчете доли принятых в университет среди мужчин и женщин можно увидеть, что поступили 46% подавших заявление мужчин и всего 30% женщин. 16% пунктов — это достаточно большая разница и маловероятно, что это просто случайное отклонение. Однако, принимая во внимание информацию о факультетах, на которые подаются заявления, различные проценты отказов показывают различную сложность поступления на факультет, и в то же время это показало, что женщины, как правило, подают заявления на более конкурентоспособные факультеты с более низкими показателями приема, даже среди квалифицированных абитуриентов (например, на английском языке). факультет), в то время как мужчины, как правило, поступали на менее конкурентоспособные факультеты с более высокими

показателями поступления (например, на инженерный факультет). Объединенные и скорректированные данные показали "небольшой, но статистически значимый перекос в пользу женщин" [3].

Парадокс Симпсона возникает, когда вы не учитываете релевантную информацию при анализе данных, например агрегируете данные и теряете важные детали в процессе.

Однако подобные явления также создают прецедент для манипуляции данными и введения в заблуждение. Подобным образом, уже в 1812 году появился термин джерримендеринг, предполагающий произвольную демаркацию избирательных округов с целью искусственного изменения соотношения политических сил в них и, как следствие, в целом на территории проведения выборов. Схематическое представление подобного явление, когда меньшинство при грамотном делении на группы окажется в выигрышном положении, на рисунке 1 [4].

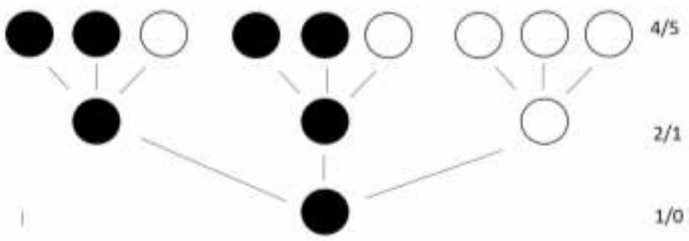

Рисунок 1 – Пример манипуляции данными с помощью джерримендеринга

Еще одним важным примером является феномен Уилла Роджерса, который заключается в том, что перемещение (численного) элемента из одного множества в другое может увеличить среднее значение обоих множеств, как показано на рисунке 2.

$$\left\{\begin{array}{lll} R & = \{1,2,3,4\} & (\text{mean} = 2.5) \\ S & = \{\mathbf{5},6,7,8,9\} & (\text{mean} = 7) \end{array}\right\} \rightarrow \left\{\begin{array}{lll} R & = \{1,2,3,4,\mathbf{5}\} & (\text{mean} = 3) \\ S & = \{6,7,8,9\} & (\text{mean} = 7.5) \end{array}\right\}$$

Рисунок 2 – Пример феномена Уилла Роджерса

Одним из реальных примеров этого феномена считается медицинский случай улучшения диагностики рака, при котором происходит повышение продолжительности жизни среди и больных, и здоровых людей, за счет того, что все подозрительные здоровые люди приписываются к больным [5].

Неправильные выводы можно сделать не только из-за неправильно проанализированных данных, но и из неграмотно представленных [6]. Например:

- изменение минимальных значений по оси OY. Чем больше минимальное значение по оси Y, тем более масштабно выглядит график.
- выбор первой и последней точки в серии, чтобы был наилучший прогресс – начать с низких показателей, закончить высокими.
- анализ на коротком интервале – оставить в выборке лишь короткий интервал, подтверждающий нашу теорию.
- выдача корреляции за причинно-следственную связь.

Осознание присутствия данных проблем позволяет разработать более эффективные методы обработки данных и анализа информации, направленные на минимизацию влияния искажений и повышение достоверности результатов. Подобный подход имеет важное значение как в академических исследованиях, так и в практических приложениях, где качество принимаемых решений непосредственно зависит от качества анализа данных.

Список использованных источников:

1. Барсегян, А.А. Методы и модели анализа данных: OLAP и Data Mining / А.А. Барсегян., М. С. Куприянов, В. В. Степаненко, И. И. Холод. — СПб.: БХВ-Петербург, 2004. — 336 с.

2. Subjective probability: A judgment of representativeness / Kahneman D, Tversky A // Cognitive Psychology, 1972 – P. 430–454

3. Sex Bias in Graduate Admissions: Data From Berkeley / P.J. Bickel, E.A. Hammel and J.W. O'Connell // Science, vol. 187, 1975

4. There's a simple way to end gerrymandering / Andrew Prokop // Wayback Machine, 2016

5. The Will Rogers phenomenon. Stage migration and new diagnostic techniques as a source of misleading statistics for survival in cancer / Feinstein AR, Sosin DM, Wells CK // The New England Journal of Medicine. 312 (25): 1604–8, 1985

6. How to lie with statistics / Huff, Darrell // New York, Norton, 1954