Анкета участника

Информация	Описание
ФИО студента	Молькин Николай Владимирович
Направление/специальность	080801.65 «Прикладная информатика (в экономике)»
Вуз	Государственное образовательное учреждение высшего профессионального образования Нижегородский государственный архитектурно-строительный университет
Вуз-партнер	Да
Город	Нижний Новгород
Кафедра	Информационные системы в экономике
ФИО зав. кафедрой	Папкова М.Д.
Тема ВКР	ПРИМЕНЕНИЕ СОВРЕМЕННЫХ ИНФОРМАЦИННЫХ ТЕХНОЛОГИЙ И ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ АНАЛИЗА ДАННЫХ В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ СПОРТИВНЫХ СОСТЯЗАНИЙ
Версия Deductor	5.2
Дата защиты	23.06.2010
Оценка	Отлично
Руководитель ВКР	Канд. ф.-м. наук, доцент Прокопенко Наталья Юрьевна
Представлено	▪ Анкета участника ▪ Аннотация ▪ Пояснительная записка ▪ Сканированные титульные листы ▪ Рецензия ▪ Отзыв научного руководителя ▪ Сценарии Deductor (с данными) ▪ Презентация Power Point

Аннотация

ПРИМЕНЕНИЕ СОВРЕМЕННЫХ ИНФОРМАЦИННЫХ ТЕХНОЛОГИЙ И ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ АНАЛИЗА ДАННЫХ В ЗАДАЧЕ ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ СПОРТИВНЫХ СОСТЯЗАНИЙ

Прогнозирование – это взгляд в будущее, оценка возможных путей развития, последствий тех или иных решений. Оценка предполагаемого исхода процесса может не только помочь в планировании, но и принести прибыль при правильном использовании информации. Наиболее ярко это можно наблюдать в букмекерских конторах, которые в наши дни нашли широкое распространение. Некоторые букмекеры уже давно экспериментируют с политическими и другими экзотическими ставками, например победа в "Евровидении", президентские выборы и т.п.

Задача прогнозирования результатов спортивных состязаний является достаточно актуальной, так как этот вопрос затрагивает различные слои общества, в том числе аналитиков букмекерских контор, при составлении линий коэффициентов для ставок, а также игроков, которые заинтересованы в получении максимального выигрыша. Об актуальности свидетельствуют следующие факты:

1. Букмекерские конторы держат целый штат аналитиков, которые занимаются прогнозированием исходов спортивных состязаний и потоков денежных средств на различные события.
2. Только лучшие игроки способны ежегодно выигрывать свыше 55 % своих ставок, при этом следует отметить, что игроки, получающие такой процент, вынуждены тратить «львиную» долю своего рабочего времени на составление прогнозов.

Таким образом, приведённые факты показывают, что автоматизация прогнозирования исходов спортивных состязаний, способна заинтересовать целевую аудиторию.

Целью настоящей работы является применение современных информационных технологий и интеллектуальных методов анализа данных в задаче прогнозирования результатов спортивных состязаний, на примере игр большого тенниса.

Для достижения поставленной цели в представленной работе решаются следующие основные задачи:

1. автоматизация процесса сбора необходимых данных;
2. изучение методов прогнозирования и анализа данных Data Mining с их дальнейшим практическим применением в выбранной области;
3. выбор наиболее оптимальной стратегии расчёта размера ставки;
4. оценка экономического эффекта от практического использования разработанной системы.

Работа содержит введение, 3 главы, заключение, библиографический список в количестве 24 наименований, 3 приложения, 45 рисунков, 8 таблиц.

В первой главе данной работы, после изучения существующих методов организации процессов проведения сбора и анализа информации по спортивным состязаниям, рассматривается проектирование основных процессов и разработка моделей процессов типовой букмекерской компании, построение инфологической и даталогической моделей, а также разработка автоматизированной информационной системы на основе этих моделей.

В данной работе рассмотрена автоматизация не всех процессов компании, а только части, а именно процесса сбора данных из сети Интернет и построения на их основе прогноза исхода спортивных состязаний. При сборе данных происходит получение статистической информации о спортивных состязаниях. Ручной сбор данных, имеющий большие временные и трудовые затраты, заменяется автоматическим, организованным на основе технологии «парсинг».

Подход к прогнозированию на основе технологий Data Mining позволяет устранить недостатки традиционных подходов с использованием рейтинговых алгоритмов (низкое качество прогноза, а также большие временные, трудовые и финансовые затраты на составление прогноза).

Вторая глава «Теоретические аспекты системы прогнозирования» посвящена вопросам рассмотрения методов анализа спортивных состязаний на предмет построения рейтингов.

Существуют несколько методов анализа данных спортивных состязаний: основанные на прогностической системе рейтингов (два наиболее часто применяемых в спорте типа рейтингов – очковая система и ЭЛО) и на основе технологий Data Mining. Очковая система обладает существенным недостатком, связанным с тем, что при ее расчете абсолютно не учитывается класс противника, с которым играет теннисист, то есть за победу над «слабым» противником, он получает столько же очков, сколько и за победу над «сильным» соперником. Данный недостаток влечет за собой слабую «прогностическую» способность очковых систем. Частично устранить недостатки очковых систем, позволяет рейтинг ЭЛО. Данная система была разработана венгерским ученым-физиком Арпадом Эло. В 1997 году Боб Руньян перенес систему Эло на область международного тенниса и разместил полученные результаты в интернете. Он стал хозяином первого веб-сайта теннисного рейтинга Эло. Факторы, которые учитываются при расчете нового рейтинга: предыдущий рейтинг игрока; значимость текущего чемпионата; разница голов в матче; результат встречи; ожидаемый исход. Этот подход обладает следующими недостатками:

- Большая зависимость от субъективного мнения эксперта, который проставляет коэффициенты в рассчитываемом рейтинге.
- Линейная зависимость между величинами, на базе которых строится рейтинг, тогда как большинство зависимостей в реальном мире нелинейны.
- Предположения о статистических свойствах разности или отношения рейтингов на практике могут не выполняться.
- Высокая стоимость адаптации рейтингов, поскольку способы расчета рейтинга, применимые для одного чемпионата, могут быть не применимы к другому чемпионату того же вида спорта и для корректировки коэффициентов, на базе которых считаются рейтинги, приходится привлекать высококвалифицированных аналитиков.

Следствием вышеперечисленных недостатков, является низкое качество прогноза на базе рейтингов, а также большие временные, трудовые и финансовые затраты на составление прогноза.

Подход к прогнозированию на основе технологий Data Mining (логистическая регрессия, деревья решений, нейронные сети) позволяет устранить недостатки традиционных подходов. Следует отметить, что данные технологии уже достаточно давно применяются к спортивному прогнозированию на Западе.

В проведенном исследовании был применен весь спектр технологий Data Mining для прогнозирования результатов Уимблдонского турнира сезона 2010 года. Разработанная система работает на статистике предыдущих игр и дает прогноз в автоматическом режиме. Кроме того, система может быть адаптирована и для других чемпионатов, для этого необходимо переобучить модели на базе статистики нового чемпионата.

Третья глава посвящена реализации информационной системы прогнозирования результатов спортивных состязаний.

Разработанная система состоит из 4-х модулей:

1. Модуль сбора, загрузки и предварительной очистки данных;
2. Хранилище данных;
3. Прогностический модуль;
4. Финансовый модуль.

Каждый из модулей выполняет свои функции и решает свои задачи.

1 Аппарат сбора, очистки и загрузки данных преобразовывает информацию из разнородных источников к форме, пригодной для анализа. В хранилище данных должна консолидироваться информация из всех доступных источников, где может иметься необходимая для анализа информация. Также производится автоматическая или полуавтоматическая корректировка ошибок в данных перед загрузкой их в хранилище. Очистка является необходимым шагом для получения качественного результата. Реальные данные очень часто содержат избыточную или некорректную информацию, которую желательно удалить или очистить до загрузки в хранилище. Кроме того, во многих случаях необходимо перед загрузкой трансформировать данные. Перед загрузкой данных в хранилище данных можно автоматически провести все необходимые действия, такие как редактирование аномалий, заполнение пропусков, удаление шумов и прочее, и загрузить в хранилище очищенные и нужным образом подготовленные данные.

Информационно-эмпирическая база данного исследования формировалась на основе данных, собранных на сайтах www.stevegtennis.com, home17.inet.tele.dk/wta/. Для сбора данных использовалась технология парсинга для необходимых разделов представленных сайтов. В практической реализации парсер был осуществлён в виде программ, написанных на php. Текст кода модуля сбора информации представлен в Приложении Б. В результате работы парсера была сформирована таблица со следующими полями: порядковый номер записи, тип покрытия, название чемпионата, дата проведения, победитель, проигравший, результат, ранг раунда, тип чемпионата. Результаты работа парсера собраны в базу данных MySql, откуда произведена выгрузка в формат Excel. Результирующая таблица содержит 28 581 записей, аккумулирующих данные за пять лет, создавая репрезентативную выборку.

2 Анализируемая информация консолидируется в специализированном хранилище данных. Хранилище данных ориентировано на решение именно задач анализа, со специфичными для этих задач механизмами хранения данных. Использование единого хранилища позволяет гарантировать непротиворечивость данных и централизованное хранение, а так же автоматически обеспечивает всю необходимую поддержку процесса анализа данных. Хранилище данных содержит специальный семантический слой, обеспечивающий возможность работы с ним пользователю без необходимости вникать в особенности хранения данных - пользователь оперирует привычными терминами – «дата», «игрок», «чемпионат».

Создание интегрированного хранилища данных «Большой теннис», а также организация обработки накопленной информации было реализовано на базе Deductor Warehouse. Так как в проекте используется учебная версия Academic Deductor Studio, то возможность прямого экспорта данных из MS Excell отсутствует. Данные загружались из текстовых файлов: выигравший игрок.txt, проигравший игрок.txt, Игры.txt, Название корта.txt, название чемпионата.txt, названия сайтов.txt, Этапы игры.txt.

В главной таблице «Игры: Большой теннис», которая представляет собой процесс, поля «Номер игры», «Дата игры», «Корт», «Чемпионат», «Выигравший», «Проигравший», «Этап», «Сайт», являются измерениями, остальные поля являются атрибутами процесса. При такой структуре ХД предполагается, что уникальность точки в пространстве определяется совокупностью измерений. Структура ХД представлена на рисунке 1.

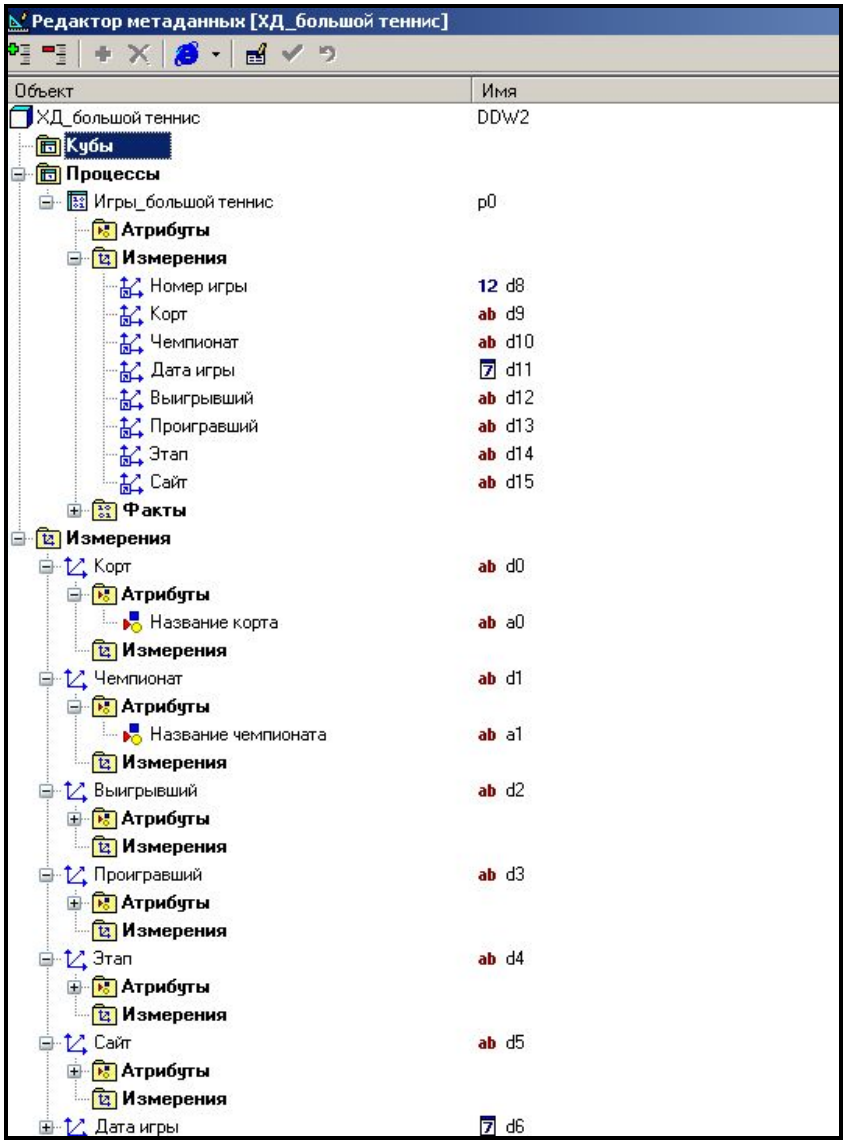

Рисунок 1 – Структура ХД в Deductor

3 Прогностический модуль разработанной информационно-аналитической системы включает несколько прогностических и классификационных моделей.

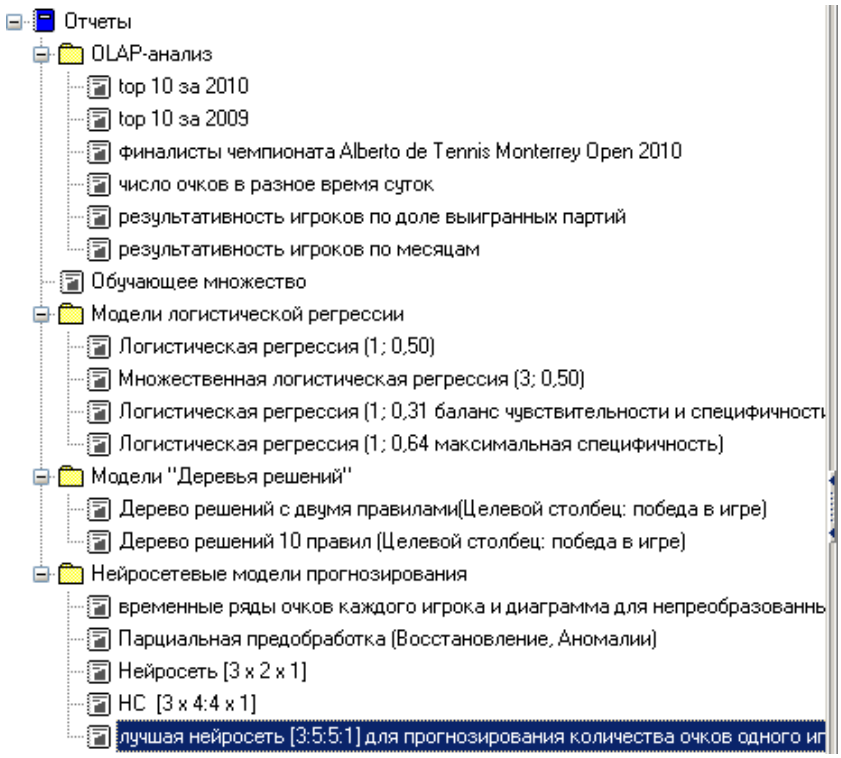

Рисунок 2 – OLAP-анализ и модели прогнозирования

Для автоматизации получения прогноза результативности каждого отдельного игрока в аналитической платформе Deductor есть две возможности: «Динамический фильтр» при выгрузке информации из ХД и обработчик «Групповая обработка». Используя «Динамический фильтр», при каждом выполнении узла импорта данных из ХД будет выводиться окно запроса, где можно будет указать необходимую информацию (например, имя игрока). Обработчик «Групповая обработка» позволяет создавать очень гибкие сценарии благодаря тому, что входной набор делится на части по указанным группам, и затем каждая группа отдельно «прогоняется» через копию цепочки узлов обработки прогностической модели.

Для того чтобы наглядно проследить динамику показателей состязаний по теннису были построены OLAP-кубы и кросс-диаграммы (см. рис. 2). Были выделены 10 лучших игроков за 2009 и 2010 года. Для ранжирования игроков был построен OLAP-куб по данным за весь период выборки, отображающий результативность игроков по доле выигранных партий (предварительно этот показатель рейтинга был получен с помощью обработчика «Калькулятор»).

Для построения моделей машинного обучения важный вопрос имеет получение качественного обучающего множества. В дипломной работе (стр. 84) приведен фрагмент сценария построения обучающего множества. Используя полученное обучающее множество и возможности аналитической платформы Deductor, были построены модели логистической регрессии, деревьев решений, а также нейросетевые модели прогнозирования (см. рис.2).

Логистическая регрессия – полезный классический инструмент для решения задачи прогнозирования вероятности победы игроков. В работе были построены модели бинарной и множественной логистической регрессии с разным порогом отсечения. Полученные в результате построения логистической регрессии таблицы содержат сведения о вероятности победы каждого из игроков (столбец «победа в игре Рейтинг»). Для анализа, интерпретации и оценки качества каждой модели логистической регрессии были получены таблицы сопряженности и ROC-кривые. Лучшей моделью оказалась модель с порогом отсечения, равным 0,64 (у нее меньшая ошибка классификации 10,73%). Таблица сопряженности этой модели показывает, что верно можно определить проигрыш игрока в

100% имеющихся примеров, а выигрыш в 13597 случаях из 19711 (69%). Анализ ROC-кривой показывает, что график находится выше диагональной линии, а численный показатель площади под кривой AUC (Area Under Curve) равен 0,886, что свидетельствует о хорошем качестве построенной модели и её пригодности для дальнейшего прогнозирования с высокой вероятностью получения достаточно точного прогноза.

Построение «Деревьев решений» стало следующим этапом моделирования. Было получено 2 модели деревьев решений (оба сформированы по 5-ти признакам, но с разным уровнем доверия), которые дают соответственно два правила (ошибка классификации составила 10,73%) и десять правил (ошибка классификации составила 0%), с помощью которых можно определять победу (1) или проигрыш (0) игрока в игре. Наиболее значимыми атрибутами оказались два атрибута: рейтинг и этап.

Следующим этапом моделирования стало построение нескольких нейросетевых моделей с разной архитектурой. Для подготовки обучающей выборки для НС, используя исходную статистику, необходимо было провести этап трансформации данных, чтобы представить данные в виде временных рядов «результативность игрока по очкам». Было проведено редактирование данных с помощью инструмента «Парциальная предобработка», в результате чего были восстановлены пропущенные значения и удалены аномалии (сглаживание). Затем была подготовлена обучающая выборка с помощью обработчика «Скользящее окно». Для прогнозирования результатов матчей отдельного игрока лучшей оказалась нейронная сеть со следующей архитектурой 3:5:5:1 : входной слой с тремя нейронами, 2 внутренних слоя с пятью нейронами и выходной слой с одним нейроном. Качество построенных нейронных сетей проверялось по диаграмме рассеивания.

Для прогнозирования результатов матчей отдельного игрока также была построена модель на основе метода декомпозиции временных рядов. Но качество этой модели нельзя считать хорошим (ошибка MAPE=55,5%).

4 Финансовый модуль. Результаты работы прогностического модуля подаются на вход финансового модуля, основное назначение которого заключается в нахождении оптимальной финансовой стратегии, то есть расчет того, сколько, когда и куда нужно ставить. В работе проведен сравнительный анализ двух стратегий: флэт-стратегии и критерия Кэли, в результате которого оптимальной по соотношению «прибыль-риск» была признана флэтовая стратегия. При проведении расчётов по данной стратегии при использовании разработанного модуля расчёта оптимального размера ставки, были получены положительные результаты.

Разработанная система прогнозирования исходов спортивных состязаний тестировалась на играх Уимблдонского турнира сезона 2010 года. Всего за 2010 год было сыграно 128 игр, но на первые два круга прогноз не делался, чтобы была возможность использовать новые данные в выборке и отследить динамику игры участников турнира. Система правильно спрогнозировала 99 игр из 128, что составляет 77,3% и является очень хорошим результатом для данной области.