МЕТОДЫ КЛАССИФИКАЦИИ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В УСЛОВИЯХ НЕОПРЕДЕЛЁННОСТИ

Дорошенко А. В., Ткаченко Р. А.
Кафедра автоматизированных систем управления, кафедра информационных технологий издательского дела, Национальный университет «Львовская политехника»
Львов, Украина
E-mail: {anastasia.doroshenko, roman.tkachenko}@gmail.com

Проанализированы особенности постановки и подходы к решению задач классификации для случаев крупноразмерных задач интеллектуального анализа данных. Представлены основы разработанных нейросетевых методов классификации, результаты проведенных экспериментов.

Введение

Многочисленные успешные примеры применений средств интеллектуальной обработки информации как в научных исследованиях, так и в различных бизнес-приложениях, привели к тому, что все больше компаний из различных отраслей хотят с помощью методов интеллектуального анализа данных добывать знания из огромных хранилищ данных, накопившихся в них благодаря развитию информационных технологий и внедрение их во все сферы человеческой деятельности. Однако, чрезвычайно большой объем хранилищ данных, используемых для поиска знаний, а также высокие требования к достоверности полученных знаний, заставляют искать новые или совершенствовать существующие методы интеллектуального анализа данных.

I. Постановка задачи

Рассмотрим задачу классификации, сформулированную организаторами ведущей немецкой лотереи South German Class Lottery. Особенностью этой лотереи является то, что количество и размер призов, что разыгрывается, определена и объявлена заранее и не зависит от того, сколько билетов продано. Продолжительность лотереи –шесть месяцев, в каждом из которых проводится отдельный розыгрыш. Принять участие в каждом из розыгрышей может только тот участник, который участвовал во всех предыдущих розыгрышах. Итак, организаторы лотереи еще до ее начала нуждаются в информации о том, сколько билетов будет продано и на какую прибыль они могут рассчитывать. Исходя из этих потребностей, была сформулирована задача - разделить всех игроков на 5 классов: те, кто принимал участие только в одном розыгрыше, но не платил за билет; те, кто принимал участие только в одном розыгрыше, но платил за билет; те, кто участвовал по крайней мере в двух розыгрышах; тот, кто принимал участие во всех розыгрышах, но не собирается играть в следующей лотереи; те, кто принимал участие во всех розыгрышах и купил хотя бы один билет следующей лотереи. Каждая ошибка классификации имеет свой вес и оценивается согласно таблице, сложившейся организаторами лотереи. Тренировочная и тестовая выборки состоят из 113 456 записей об игроках, каждый из которых имеет 70 атрибутов, в том числе: пол, возраст и семейное положение игрока, информация о его банк и марку автомобиля, кредитную привлекательность и тому подобное.

Сложность такой задачи заключается в том, что из неопределенности, вызванные пробелами в данных, их противоречивостью и т.д., для нее не выполняется гипотеза компактности, что лежит в основе многих методов классификации [1], различные классы перекрываются между собой, что делает невозможным их разделение гиперповерхности простого вида. Для решения этой проблемы предлагается применить кусковой метод построения разделяющих поверхностей на основе модели геометрических преобразований [2], модифицированный для задачи классификации на более чем два класса, который, с одной стороны, позволяет учесть нелинейность задач добычи данных, но не требует большого количества времени для выполнения. С использованием дерева разделения на классы можно объединять в отдельные кластеры векторы данных, имеющих сходные входные показатели и анализировать их независимо друг от друга. После того, как получено значение штрафных баллов по каждому из кластеров, они суммируются. Такой подход позволяет существенно повысить общую точность классификации.

II. Классификация с использованием метода штрафов и поощрений

Также, для учета различного веса ошибок, для каждого из кластеров используем метод штрафов и поощрений [3]. Соответственно, в данном случае задача классификации сводится к задаче максимизации суммы поощрительных баллов. Таким образом, предлагаемый нами алгоритм сочетание использования модели геометрических преобразований с методом штрафов и поощрений имеет вид: Алгоритм классификации с использованием метода штрафов и поощрений

1. В обучающей выборке заменяем идентификаторы классов соответствующими коэффициентами:
   Класс 1 -> $(a_{11}; a_{12}...; a_{1k})$
   Класс 2 -> $(a_{21}; a_{22}...; a_{2k})$
   ...
   Класс K -> $(a_{k1}; a_{k2}...; a_{kk})$;
2. На полученной обучающей выборке учим нейроподобную структуру на основе модели геометрических преобразований (МГП).
3. Через обученную нейронную сеть пропускаем тестовые данные.
4. Анализируем коэффициенты$(a_{11}; a_{12}...; a_{1k})$, полученные на выходах МГП для каждого вектора входных данных из тестового файла по правилу «победитель забирает все».
5. Для тестовой выборки подсчитываем количество штрафных баллов в соответствии с матрицей штрафов.
6. Основной целью алгоритма является максимизация поощрительных баллов.

Как правило, предварительно определяют или сумму штрафов, которая является приемлемой для данной задачи, или время, в течение которого будет выполняться минимизация - это условия остановки выполнения алгоритма. Рассмотрим результаты экспериментов, проведенных для сформулированной выше задачи.

Таблица 1 – Результаты классификации с помощью нейроподобной структуры МГП

	Сумма баллов до кластеризации	Сумма баллов после кластеризации
Обучение	62745	67515
Тестирование	11915	13615

III. Совершенствование метода штрафов и поощрений путем применения модифицированного алгоритма имитации отжига металла

Предлагаем рассмотреть сочетание метода кусковой построения разделяющих поверхностей на основе модели геометрических преобразований [3] и метода глобальной оптимизации - алгоритма имитации отжига металла [4]. На рис.1. изображена структурная схема разработанной нейроподобной структуры на основе МГП, где $(x_1, x_2, ..., x_n)$ - первичные признаки объектов классификации - входные данные, $(GK_1, GK_2, ..., GK_n)$ - главные компоненты, полученные на основе входных данных ,$(w_1, w_2, ..., w_n)$ - весовые коэффициенты, y - выход, задающий принадлежность к определенным классам. Функционирование такой нейроподобной структуры можно описать формулой $y = \sum_{i=1}^{n}(GK * w_i)$. Метод имитации отжига металла предлагается применять для оптимизации весовых коэффициентов так, чтобы результирующая сумма поощрительных баллов была максимальной.

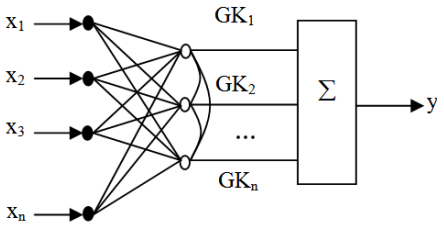

Рис. 1 – Структурная схема нейроподобной структуры на основе МГП

Модифицированный алгоритм имитации отжига металла объединены с методом штрафов и поощрений

1. Запустить процесс с начальной точки, выбранной случайно при заданной начальной температуре $T = T_{max}$, равной минимальному значению поощрительных баллов в начальной точке.
2. Пока, повторить L = 100 раз такие действия
   - выбрать новое решение $w^{'}$ с окрестности w;
   - рассчитать изменение целевой функции, где значением целевой функции $\Delta = E(w^{'}) - E(w)$ является сумма поощрительных баллов;
   - если $\Delta \leq 0$- принять $w = w^{'}$; иначе, при $\Delta > 0$, принять $w = w^{'}$ с вероятностью $exp^{\frac{-\Delta}{T}}$, путем генерации случайного числа R из интервала (0,1) с последующим сравнением его со значением $exp^{\frac{-\Delta}{T}}$; если $exp^{\frac{-\Delta}{T}} >$ R, принять новое решение $w = w^{'}$; в противном случае - проигнорировать его.
3. Уменьшить температуру $T = r * T$ с использованием коэффициента уменьшения r, выбраным из интервала (0,1) и вернуться к пункту 2. Предлагается использовать значение $r = 0,9$ Разработан модифицированный алгоритм имитации отжига металла соединяются с методом штрафов и поощрений применяется для улучшения результатов классификации для каждого из кластеров.

IV. Список литературы

1. Васильев В.И., Коноваленко В.В., Горелов Ю.И. Имитационное управление неопределенными объектами. - К.: “Наукова думка”, 1989. - 216с.
2. Дорошенко А.В. Нейромережний розв’язок задач класифікації в умовах неповноти інформаційного базису // Моделювання та керування станом еколого-економічних систем регіону: Зб.наук.пр. - Вип.3. - Київ, 2006. - С. 115-122.
3. Tkachenko R., Tkachenko P., Tkachenko O.,Schmitz J. Geometrical Data Modelling // Збірник матеріалів міжнародної наукової конференції ”Інтелектуальні системи прийняття рішень та прикладні аспекти інформаційних технологій” (ISDMIT’ 2006). - Т.2. - С.279-283.
4. Осовский С. Нейронные сети для обработки информации / Пер. с польского И.Д. Рудинского. - М.: Финансы и статистика, 2004. - 344с.