Использование нелинейных преобразований признаков для поиска скрытых закономерностей в данных

Е.Н. Згуральская[1]

[1]*Ульяновский технический университет, Созидателей, 13а, Ульяновск, Россия, 432072*

Аннотация

Для обнаружения скрытых закономерностей при анализе данных используются нелинейные преобразования на основе вычисления значений функции принадлежности к классам по каждому признаку. Проверяется истинность утверждения (гипотезы) о наличии наборов признаков, по которым объекты классов по обобщённым оценкам будут без ошибок (корректно) разделены на числовой оси.

Ключевые слова

Нелинейные преобразования, обобщённые оценки объектов, интервальные методы

1. Введение

Одним из средств для эффективного решения задач машинного обучения является использование алгоритмов преобразования данных. Разрабатываются новые алгоритмы распознавания без каких–либо предположений о природе среды данных (как того требуют традиционные статистические методы), методы для анализа информации даже в условиях нелинейной зависимости признаков [1]. Доказательством этому служит предложенная в [2] методика отбора информативных наборов признаков для линейного дискриминанта Фишера с использованием отношения внутриклассового сходства и межклассового различия, определяемого через функцию Лагранжа, критерия для вычисления оптимальной границы между объектами из разных классов. Показано, что применение оптимальной границы в качестве порога линейной решающей функции увеличивает обобщающую способность алгоритма при распознавании.

2. Постановка задачи и методы решения

Одной из форм применения интервальных методов является преобразование шкал измерений. В [3] описан критерий для преобразования значений количественных признаков в градации номинальных. шкале измерений. Такое признаковое пространство в номинальной шкале измерений будем называть сырым.

Нелинейное преобразование номинального признака в [4] сводится к замене значений его градаций на значения функции принадлежности объектов к классам и вычисление границы между классами. Значения границы используется для представления описаний объектов в {1,2} по каждому признаку. Такое пространство будем называть унифицированным.

Рассматривается множество объектов $E_0=\{S_1,\ldots,S_m\}$, разделенное на два непересекающихся класса K_1 и K_2. Каждый объект $S_u \in E_0$, $u=1,\ldots,m$ описываются набором разнотипных признаков $X(n)=(x_1,\ldots,x_n)$, $\delta(\delta \le n)$ из которых измеряются в номинальной, $n-\delta$ в интервальных шкалах.

В предлагаемых алгоритмах распознавания в процессе обучения используется обобщенные оценки объектов [3]. Многообразия значений обобщённых оценок формируется в зависимости от выбора различных вариантов начальных приближений для стохастического алгоритма оптимизации. Потребность в выборе начальных приближений отпадает если описания объектов представлены значениями градаций номинальных признаков. Добиться этого можно либо исключив количественные признаки из описания, либо применив отображение их значений в градации номинальных.

Для номинальных признаков определен перевод в унифицированную форму для описания объектов через бинарную матрицу.

Считается, что на наборе признаков $X(n)$ определена процедура вычисления значений функции принадлежности объектов к классам. Требуется:

– получить представление градаций номинальных признаков в интервальной шкале измерений в виде значений функции принадлежности объектов к классам;

– определить порог (границу) между классами для разбиения признаков на непересекающиеся интервалы по максимуму произведения внутриклассового сходства и межклассового различия;

– для всех признаков (номинальных и количественных) реализовать преобразование их значений в {1,2};

– вычислить обобщённые оценки объектов по их описанию в {1,2} на заданных наборах признаков.

Проводился эксперимент на выборке данных из [4, 5]. Доказана эффективность нелинейного преобразования по каждому признаку с использованием функций принадлежности объектов к классам. Точность распознавания по обобщённым оценкам на наборах сырых и унифицированных признаков была соответственно 98,33% и 100%.

3. Заключение

В рамках теории распознавания образов описанное нелинейное преобразование признаков рассматривается как процесс формирования базовых алгоритмов ансамбля алгоритмов метамодели. Сравнение показателей точности служат доказательством эффективности использования ансамбля алгоритмов метамодели с использованием нелинейных преобразований признаков.

4. Литература

[1] Кузнецова, А.В. Возможности использования методов Data Mining при медико-лабораторных исследованиях для выявления закономерностей в массивах данных / А.В. Кузнецова, О.В. Сенько // Врач и информационные технологии. – 2005. – № 2. – С. 49-56.

[2] Игнатьев, Н.А. Анализ данных и принятие решений с помощью логических закономерностей в форме полуплоскостей / Н.А. Игнатьев, Д.Ю. Саидов // Известия Самарского научного центра РАН. – 2017. – Т. 19, № 4(2). – С. 294-299.

[3] Игнатьев, Н.А. Вычисление обобщённых показателей и интеллектуальный анализ данных / Н.А. Игнатьев // Автоматика и телемеханика. – 2011. – № 5. – С. 183-190.

[4] Игнатьев, Н.А. Поиск скрытых закономерностей, влияющих на общую выживаемость больных, методами интеллектуального анализа данных / Н.А. Игнатьев, Е.Н. Згуральская, М.В. Марковцева // Искусственный интеллект и принятие решений. – 2020. – № 3. – С. 73-80.

[5] Ignatyev, N. Nonlinear transformation of signs and the search for patterns in the data of patients with chronic lymphocytic leukemia / N. Ignatyev, E. Zguralskaya, M. Markovtseva // CEUR Workshop Proceedings. – 2020. – № 2667. – P. 333-336.