УДК 519.23

В.А. Алексеева

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ БИНАРНОЙ КЛАССИФИКАЦИИ

Алексеева Венера Арифзяновна, *кандидат технических наук, окончила экономико-математический факультет Ульяновского государственного технического университета. Доцент кафедры «Прикладная математика и информатика» Ульяновского государственного технического университета. Имеет научные труды в области статистических методов. [e-mail: v.a.alekseeva@bk.ru].*

Аннотация

В статье рассматривается задача бинарной классификации объектов, для решения которой предлагается использование методов машинного обучения. Машинное обучение – подраздел искусственного интеллекта, математическая дисциплина, использующая разделы математической статистики, численных методов оптимизации, теории вероятностей, дискретного анализа. Целью машинного обучения является частичная или полная автоматизация решения сложных профессиональных задач в самых разных областях человеческой деятельности, таких как обнаружение объектов, распознавание речи, образов, медицинская диагностика, диагностика технических объектов и т.д.

В статье для бинарной классификации объектов предлагается использовать следующие методы: деревья решений, нейронные сети, дискриминантный анализ, байесовский классификатор, метод опорных векторов, логистическая регрессия, бэггинг деревьев решений, метод эмпирической функции (МЭФ) и нечеткий логический вывод на базе модели Сугэно. Эффективность классификации оценивается с помощью ряда характеристик: среднеквадратической ошибки, ROC-кривой, показателя AUC и т.д.

Для повышения точности прогнозирования классов объектов предлагается провести сравнительный анализ эффективности рассматриваемых методов при различных порогах отсечения. Также предлагается использование комбинации моделей, так называемого агрегированного классификатора.

Ключевые слова: бинарная классификация, машинное обучение, агрегированный классификатор, порог отсечения.

THE USE OF MACHINE LEARNING METHODS FOR BINARY CLASSIFICATION

Venera Arifzianovna Alekseeva, *Candidate of Engineering; graduated from the Faculty of Economics and Mathematics of Ulyanovsk State Technical University; Associate Professor at the Department of Applied Mathematics and Informatics at UlSTU; an author of scientific papers in the field of statistical methods. e-mail: v.a.alekseeva@bk.ru.*

Abstact

The article deals with the problem of objects binary classification. To solve this problem, the use of machine learning methods should be provided. Machine learning is a subsection of artificial intelligence. It is a mathematical discipline using subsections of mathematical statistics, numerical optimization methods, probability theory, and discrete analysis. The goal of machine learning is a partial or complete automation of solutions of complex professional tasks in different fields of human activity such as speech recognition, image recognition, medical diagnostics, diagnostics of technical facilities etc.

The use of the following methods for binary classification: decision trees, neural networks, nearest-neighbor method, discriminant analysis, Bayesian classifier, a support vector machine, logistic regression, decision trees begging, method of the empirical function and fuzzy inference based on the basis of Sugeno model are proposed. Classification efficiency is assessed with the use of a number of characteristics: mean-square error, ROC-curve, AUC index, etc.

In order to improve the accuracy of classes objects prediction, a comparative analysis of the effectiveness of these methods at various cut-offs and a combination of models (so-called aggregate classifier) are offered.

Keywords: binary classification, machine learning, aggregate classifier, cut-off.

Введение

Рассмотрим задачу бинарной классификации объектов, в которой каждый объект $K_i (i = 1, ..., N)$ характеризуется m-мерным вектором признаков $(X_1 ... X_m)$. Эти характеристики (или признаки) могут принимать как числовые, так и нечисловые значения и образуют выборку для дальнейших исследований. Требуется на основании значений признаков предсказать выходную характеристику объектов y, принимающую два значения (будем обозначать их 0 и 1).

В качестве примеров таких задач могут служить задачи распознавания технического состояния объекта и классификации этого состояния (нормального или аномального), обнаружения факта передачи или отсутствия единственного сигнала, нормального или патологического состояния органа и т.д.

При решении задач классификации широкое распространение получили методы машинного обучения [2], такие как деревья решений, нейронные сети, дискриминантный анализ, байесовский классификатор, метод опорных векторов, логистическая регрессия и др.

1 Методы бинарной классификации объектов

Рассмотрим основные методы бинарной классификации объектов [2].

Деревья принятия решений [1]. Значения выходной бинарной переменной предсказываются на основе характеристик объекта с помощью построения дерева. На каждом этапе разделение на классы производится по самому значимому фактору или другим критериям.

Нейронные сети [7]. На обучающей выборке производится обучение сети. После обучения сеть способна определять, к какому классу относится входной сигнал.

Дискриминантный анализ [5]. Этот метод заключается в построении системы линейных регрессионных моделей для каждого класса объекта.

Байесовский классификатор [3]. В основе байесовского подхода лежит принцип максимального использования априорной информации об объектах. При этом применяется теорема Байеса.

Метод опорных векторов [9]. Основная суть метода заключается в переводе исходных векторов в пространство более высокой размерности, а также в поиске оптимальной гиперплоскости, разделяющей классы клиентов наилучшим образом.

Логистическая регрессия [4]. Применяется для предсказания вероятности возникновения некоторого события по значениям множества признаков. Если предсказанное значение вероятности больше 0,5, то считаем, что $y = 1$, в противном случае $y = 0$.

Бэггинг (bagging) деревьев решений [9]. Для получения решения задачи классификации может быть использован метод, относящий объект в тот класс, куда его отнесло большинство алгоритмов. Процедура бэггинга показывает высокий прирост обобщающей способности по сравнению с алгоритмом, обученным с помощью базового метода по исходной обучающей выборке, в тех случаях, когда вариационная составляющая ошибки базового метода высока. К таким моделям относятся, в частности, решающие деревья и нейросетевые методы. При использовании в качестве базового метода решающих деревьев процедура бэггинга приводит к построению ансамблей решающих деревьев (решающих лесов).

Нечеткая логика. Классификация объектов может проводиться с применением концепции нечеткой логики. Наиболее популярными моделями нечеткого вывода являются модели двух типов: Мамдани и Сугэно [8]. Они отличаются форматом базы знаний и процедурой дефазификации.

В модели типа Сугэно нечеткая база знаний имеет вид:

$$\bigcup_{p=1}^{k_j} \bigcap_{i=1}^{n} \left(x_i = a_{i,jp}\right) \rightarrow y = b_{j,0} + b_{j,1} x_1 + ... + b_{j,n} x_n, j = \overline{1,m}, \quad (1)$$

где $a_{i,jp}$ – лингвистический терм, которым оценивается переменная x_i в строке с номером jp;

kj – количество строк-конъюнкций, в которых выход y оценивается лингвистическим термом dj;

m – количество термов, используемых для лингвистической оценки выходной переменной y;

bj, i – некоторые числа.

Правила при этом задаются не нечеткими термами, а линейной функцией от входов. Таким образом, можно сказать, что подобная база знаний является гибридной, так как ее правила в левой части содержат нечеткие множества, а в правой – заключения в виде четкой линейной функции. Результирующее значение выхода y определяется как суперпозиция линейных зависимостей в данной точке n-мерного пространства. Это может быть взвешенное среднее или взвешенная сумма.

Метод эмпирической функции. Начинаем просматривать исходную матрицу данных по столбцам. Пронумеровав все возможные значения каждой из компонент столбца по отдельности и проделав эту процедуру для всех столбцов, можно записать любую из возможных реализаций каждого вектора X в виде N-разрядного числа $(x_1, x_2, ..., x_N)$ – кода кластера в N-мерном пространстве, соответствующего данной реализации вектора X. При этом предполагается, что количество возможных значений каждой из компонент не превышает десяти. В противном случае разрядность числа увеличится, однако это не влияет на ход дальнейших рассуждений. На выходе получаем закодированную матрицу. Следующим этапом алгоритма является поиск одинаковых строк в уже закодированной исходной матрице, так как строка закодированной матрицы по сути является кодом кластера в N-мерном пространстве, соответствующим данной реализации вектора X, а кластеры не должны повторяться. Применение такой модели состоит в том, что по признакам нового объекта

можно определить номер и код кластера, к которому он относится, и из таблицы извлекается соответствующее значение МЭФ.

Агрегированный классификатор. Для решения задачи прогнозирования класса исследуемого объекта в настоящее время существует множество методов и моделей, каждые имеют свои преимущества и недостатки. Например, нельзя использовать МЭФ для прогнозирования данных, набор значений признаков которых не совпадает хотя бы с одним набором из обучающей выборки, а для применения байесовского подхода необходимо прежде привести исходные данные к интервальной шкале, чтобы переменные были дискретными, иначе это может привести к потере значимой информации. Нет универсальной модели, с помощью которой можно было бы с высокой точностью оценить принадлежность объекта к тому или иному классу.

Для повышения эффективности задачи классификации возможно использование комбинации моделей, то есть нескольких методов классификации одновременно.

Существуют подходы (бэггинг и бустинг), в которых используется один и тот же классификатор (например, деревья решений), построенный на разных частях обучающей выборки. Однако в зависимости от конкретной ситуации наилучшим с точки зрения точности прогнозирования может оказаться любой из методов машинного обучения.

Таким образом, представляет интерес совместное использование различных классификаторов, построенных на разных частях обучающей выборки.

Рассмотрим набор из девяти базовых перечисленных выше методов машинного обучения. Используя метод полного перебора, можно получить всевозможные комбинации различных моделей. Общее число таких комбинаций будет равно $2^9 - 9 - 1 = 502$. В общем случае при количестве исходных методов n число комбинаций равно $2^n - n - 1$.

Комбинация методов может формироваться последовательным или параллельным их соединением. При последовательном соединении результат каждого последующего классификатора зависит от предыдущего; при этом возникает проблема, связанная с порядком следования друг за другом классификаторов. Поэтому в настоящей работе использовалось параллельное соединение выбранных классификаторов.

Для формирования единого решения о принадлежности к одному из классов на основе решений отдельных методов классификации возможно агрегирование результатов по трем признакам:

- по среднему значению (вероятность принадлежности заданного объекта классу $y = 1$ определяется как среднее арифметическое значений вероятностей принадлежности объекта классу $y = 1$, найденных по всем базовым классификаторам);
- по медиане (сначала ранжируется ряд, содержащий результаты базовых классификаторов в комбинации, вероятность находится путем вычисления результата срединного классификатора в случае их нечетного количества или полусуммы результатов срединных базовых классификаторов в четном случае);
- с помощью процедуры голосования (результат агрегированного классификатора по голосованию представляет собой среднее значение результатов базовых классификаторов, которые определили факт принадлежности заданного объекта классу $y = 1$ с вероятностью $\geq 0{,}1$).

2 Программа, реализующая классификацию объектов

Описанный подход к бинарной классификации был реализован в виде программного продукта, позволяющего прогнозировать класс рассматриваемого объекта на основе обучающей выборки. Программный комплекс был разработан в среде программирования Matlab R2014a, содержащей все методы обработки исходных данных и большинство алгоритмов машинного обучения, необходимых для решения задачи классификации.

Программа позволяет проводить предварительную обработку исходных данных: восстановление пропущенных данных; дискретизацию характеристик; кодирование нечисловых данных; выбор статистически значимых признаков.

В программе реализованы все рассмотренные выше методы классификации, а также агрегированный классификатор с возможностью выбора критерия агрегирования (по среднему значению, по медиане или с помощью процедуры голосования). Имеется возможность включения в программу новых методов классификации.

При построении классификатора на обучающей выборке есть опасность «подгонки» под тренировочные данные. Для получения несмещенных оценок качества классификаторов применяется метод 10-кратной перекрестной проверки, заключающийся в разделении исходной выборки на 10 непересекающихся частей, приблизительно равных по объему. Далее в порядке очереди каждая часть выступает в роли контрольной выборки, а остальные части объединяются в обучающую выборку. Качество классификатора определяется усреднением ошибок по всем контрольным выборкам.

В результате работы программы формируются результаты для трех порогов отсечения (значение, выше которого объект признается принадлежащим классу $y = 1$): порог отсечения 0,5; оптимальный порог отсечения; порог отсечения, заданный пользователем. Оптимальным порогом классификации было выбрано наименьшее отклонение между ошибками первого рода и второго рода. Ошибка первого рода возникает, когда интересующее нас событие ошибочно не обнаружилось. Ошибка второго рода возникает, когда при отсутствии события ошибочно выносится решение о его присутствии.

Оценка качества построенных классификаторов производится с помощью следующих критериев [10]: ошибки первого и второго рода, ROC-кривые, показатель AUC, среднеквадратическая ошибка прогнозирования (MSE).

По заданным критериям пользователь может определить, какой метод или комбинация методов дают оптимальный результат для исследуемых объектов, и построить прогноз для исходного набора значений признаков. Работа агрегированного классификатора производится программой, т.е. по ряду критериев автоматически формируется оптимальная комбинация методов, после чего пользователь может сравнить результаты агрегированного классификатора с базовыми методами классификации.

Рассмотрим для примера работу программы по реализации агрегированного классификатора. В качестве исходных данных была рассмотрена выборка по заемщикам немецкого банка (http://www.stat.uni-muenchen.de/service/datenarchiv /kredit/kredit_e.html), включающая 1000 клиентов, описанных 20 признаками (статус текущего чекового счета, кредитная история, цель кредита, срок кредита, сумма кредита, средний баланс на накопительном счете, стаж работы на последнем месте, доход в %, семейное положение, поручители, постоянное проживание на последнем месте, данные об имуществе, возраст, имеющиеся кредиты, вид жилья, количество предыдущих кредитов в этом банке, вид деятельности, количество иждивенцев, наличие телефона, гражданство) и одной зависимой бинарной переменной (кредитоспособен или некредитоспособен клиент). Проведена классификация выборки с помощью всех рассмотренных выше методов, включая агрегированный классификатор. Выполнена предварительная обработка данных, включающая в себя дискретизацию ряда признаков и кодирование нечисловых данных, таких как гражданство клиентов, образование, семейное положение и т.д. Проанализированы все девять отдельных методов классификации и агрегированный классификатор. Агрегирование проводилось по среднему значению.

Для выборки был получен оптимальный из всех возможных агрегированный классификатор (порог отсечения 0,5), состоящий из следующих методов: метод опорных векторов, логистическая регрессия, бэггинг деревьев решений, МЭФ и метод нечеткого логического вывода. Результаты работы программы представлены в таблице 1. Из таблицы видно, что для данной выборки лучший результат классификации получен с помощью агрегированного классификатора:

1) среднеквадратическая ошибка меньше, чем у остальных методов;

2) самый высокий процент верных прогнозов кредитоспособных клиентов наблюдается для двух методов: агрегированного классификатора и бэггинга деревьев решений, но при этом ошибка первого рода у агрегированного классификатора ниже;

3) по некредитоспособным клиентам агрегированный классификатор дает средний результат по прогнозу, но с минимальной ошибкой второго рода.

Результаты классификации можно также представить в виде диаграммы, отображающей площади под ROC-кривыми (AUC) (рис. 1).

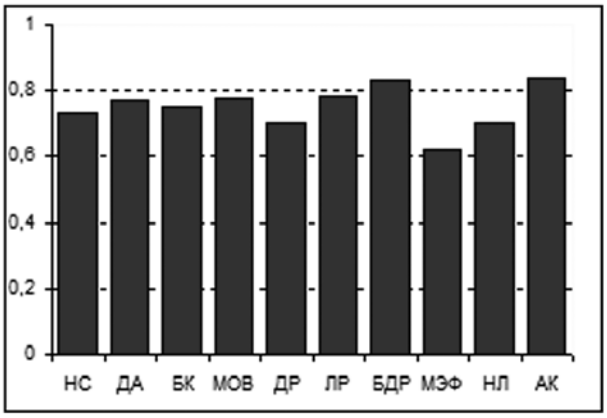

Рис.1. Площади под ROC-кривыми для первой выборки

Таблица 1.

Результаты классификации по заемщикам немецкого банка

Классификатор	Среднекв. ошибка (MSE)	Кредитоспособные $(y=1)$		Некредитоспособные $(y=0)$	
		Верный прогноз, %	Ошибка первого рода, %	Верный прогноз, %	Ошибка второго рода, %
Нейронная сеть (НС)	0,1928	85,2	56,8	43,2	14,8
Дискриминантный анализ (ДА)	0,1832	84,4	48,0	52,0	15,6
Байесовский классификатор (БК)	0,2139	73,4	36,6	63,4	26,6
Метод опорных векторов (МОВ)	0,1744	86,9	49,8	50,2	13,1
Деревья решений (ДР)	0,2458	78,6	45,4	54,6	21,4
Логистическая регрессия (ЛР)	0,1754	87,7	50,2	49,8	12,3
Бэггинг деревьев решений (БДР)	0,1532	89,0	48,9	51,1	11,0
МЭФ	0,4893	29,0	2,6	97,4	71,0
Нечеткая логика (НЛ)	0,2397	76,5	38,8	61,2	23,5
Агрегированный классификатор (АК)	0,1449	89,0	38,8	61,2	11,0

ROC-кривая [10], также известная как кривая ошибок, отображает соотношение между долей верных положительных классификаций от общего числа положительных классификаций (true positive rate) и долей ошибочных положительных классификаций от общего числа отрицательных классификаций (false positive rate) при варьировании порога решающего правила. Показатель AUC (площадь под ROC-кривой) дает количественную интерпретацию ROC-кривой. Считается, что чем выше показатель AUC, тем качественнее классификатор. Оказалось, что наиболее точный результат классификации дает агрегированный классификатор.

Представленный пример показывает возможности разработанной программы. Из всей совокупности методов классификации выбирается тот, который позволяет с наиболее высокой точностью прогнозировать кредитоспособность и некредитоспособность клиентов одновременно, при этом минимизируя среднеквадратическую ошибку и ошибки первого и второго рода. Используя выбранный в программе метод, в дальнейшем можно по заданному набору значений факторов определить, к какому классу относится исследуемый объект.

Разработанный программный продукт можно использовать для любой задачи бинарной классификации, в частности, для прогнозирования технического состояния объектов или факта обнаружения или отсутствия сигналов. Например, с помощью методов классификации можно определить, будет работоспособен или неработоспособен механический узел машины в зависимости от ряда факторов.

Заключение

Для решения задачи бинарной классификации объектов предложено использование различных методов машинного обучения, а также их комбинаций (агрегированного классификатора). Говорить об эффективности какого-либо из рассмотренных методов нецелесообразно, так как для разных выборок, даже для разных частей одной выборки, мы можем получить различные результаты. Можно только ввести ряд ограничений на выборку, связанных с особенностью того или иного метода или области исследования, например:

1) логистическая модель является чувствительной к корреляции между факторами, поэтому в модели недопустимо наличие сильно коррелированных входных переменных;

2) метод опорных векторов чувствителен к шумам и стандартизации данных;

3) для применения байесовского подхода необходимо прежде привести исходные данные к интервальной шкале, чтобы переменные были дискретными, иначе это может привести к потере значимой информации и т.д.

В результате разработанная программа подберёт наилучшую (в смысле заданных критериев) комбинацию из задействованных методов, то есть оптимальный агрегированный классификатор.

В перспективе рассматривается возможность расширения функционала программы за счет добавления новых наиболее эффективных методов классификации.

СПИСОК ЛИТЕРАТУРЫ

1. Breiman L. Random Forests // Machine Learning, 2001. – 45(1). – pp. 5–32.

2. Алексеева В.А. Использование методов интеллектуального анализа в задачах бинарной классификации // Известия Самарского научного центра Российской академии наук. – 2014. – Т. 16, №6(2). – С. 354–356.

3. Бидюк П.И., Терентьев А.Н. Построение и методы обучения байесовских сетей // Информатика и кибернетика. – 2004. – № 2. – С. 140–154.

4. Васильев Н.П. Опыт расчета параметров логистической регрессии методом Ньютона-Рафсона для оценки зимостойкости растений // Математическая биология и биоинформатика. – 2011. – Т. 6, № 2. – С.190–199.

5. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: учебник. – М. : Финансы и статистика, 2003. – 352 с.

6. Клячкин В.Н., Донцова Ю.С. Сравнительный анализ точности нелинейных моделей при прогнозировании состояния системы на основе марковской цепи // Известия Самарского научного центра Российской академии наук. – 2013. – Т. 15, № 4(4). – С. 924–927.

7. Ясницкий Л.Н. Введение в искусственный интеллект. – М. : Издательский центр «Академия», 2005. – 176 с.

8. Штовба С.Д. Идентификация нелинейных зависимостей с помощью нечеткого логического вывода в системе Matlab // Научно-практический журнал Exponenta Pro: математика в приложениях. – 2003. – №2(2). – С. 9–15.

9. Шунина Ю.С., Алексеева В.А., Клячкин В.Н. Прогнозирование кредитоспособности клиентов на основе методов машинного обучения // Финансы и кредит. – 2015. – №27(651). – С. 2–12.

10. Шунина Ю.С., Алексеева В.А., Клячкин В.Н. Критерии качества работы классификаторов // Вестник Ульяновского государственного технического университета. – 2015. – № 2(70). – С. 67–70.

REFERENCES

1. Breiman L. Random Forests. *Machine Learning,* 2001, no. 45(1), pp. 5–32.

2. Alekseeva V.A. Ispolzovanie metodov intellektualnogo analiza v zadachakh binarnoi klassifikatsii [Using of Mining Techniques in Problems on Binary Classification]. *Izvestiia Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk* [Scientific Journal of Proceedings of the Samara Scientific Center of the Russian Academy of Sciences], 2014, vol. 16, no. 6 (2), pp. 354–356.

3. Bidyuk, P. I., Terentiev A. N. Postroenie i metody obucheniia baiesovskikh setei [Bayesian Networks Construction and Learning Methods]. *Informatika i kibernetika* [Informatics and Cibernetics], 2004, no. 2, pp. 140–154.

4. Vasiliev N. P. Opyt rascheta parametrov logisticheskoi regressii metodom Nutona-Rafsona dlia otsenki zimostoikosti rastenii [Experience of Logistic Regression Parameters Calculation by Newton-Rafson Method to Estimation Resistance to Cold of Plants]. *Matematicheskaia biologiia i bioinformatika* [Mathematical Biology and Bioinformatics], 2011, vol. 6, no. 2, pp. 190–199.

5. Dubrov A. M., Mkhitaryan V. S., Troshin L. I. *Mnogomernye statisticheskie metody: Uchebnik* [Multidimensional Statistical Method. Textbook]. Moscow, Finansy i statistika Publ., 2003. 352 p.

6. Klyachkin V.N., Dontsova J.S. Sravnitelnyi analiz tochnosti nelineinykh modelei pri prognozirovanii sostoianiia sistemy na osnove markovskoi tsepi [The Comparative Analysis of Accuracy of Nonlinear Models at Forecasting of the Condition of System on the Basis of Markovsky Chain]. *Izvestiia Samarskogo nauchnogo tsentra Rossiiskoi akademii nauk* [Scientific Journal of Proceedings of the Samara Scientific Center of the Russian Academy of Sciences], 2013, vol. 15, no. 4 (4), pp. 924–927.

7. Yasnitskii L.N. *Vvedenie v iskusstvennyi intellekt* [Introduction to Artificial Intelligence]. Moscow, Izdatelskii tsentr Akademiia Publ., 2005. 176 p.

8. Shtovba S. D. Identifikatsiia nelineinykh zavisimostei s pomoshchiu nechetkogo logicheskogo vyvoda v sisteme Matlab [Identification of Nonlinear Dependences with Nonliner Logic Inference in Matlab]. *Nauchno-prakticheskii zhurnal Exponenta Pro: matematika v prilozheniiakh* [Scientific Journal Exponenta Pro: Mathematics in Applications], 2003, no. 2 (2), pp. 9–15.

9. Shunina Yu.S., Alekseeva V.A., Klyachkin V.N. Prognozirovanie kreditosposobnosti klientov na osnove metodov mashinnogo obucheniia [Forecasting the Customers' Creditworthiness through Machine Learning Methods]. *Finansy i kredit* [Finance and Credit], 2015, no. 27 (651), pp. 2–12.

10. Shunina Yu.S., Alekseeva V.A., Klyachkin V.N. Kriterii kachestva raboty klassifikatorov [Classifier Performance Criteria]. *Vestnik Ulyanovskogo gosudarstvennogo tekhnicheskogo universiteta* [Bulletin of Ulyanovsk State Technical University], no. 2 (70), 2015, pp. 67–70.