Разработка ансамбля алгоритмов классификации с использованием энтропийного показателя качества для решения задачи поведенческого скоринга

© И. А. Кузнецов © В. С. Киреев

Национальный исследовательский ядерный университет «МИФИ»,
Москва, Российская Федерация
iakuznetsov@mephi.ru vskireev@mephi.ru

Аннотация

С увеличением объёма цифровой информации в мире возрастает актуальность задачи фильтрации и обработки таких данных. С целью выявления действительно необходимой и полезной информации для пользователя, применяются подходы, основанные на принципах классификации объектов и отнесения исходного объекта к группе наиболее похожих на него. Основой для классификации выступают алгоритмы машинного обучения, а сама классификации успешно применяется в различных областях интенсивного использования данных, в частности, в рекомендательных системах. Представленная статья посвящена описанию разработанного ансамбля алгоритмов классификации при построении рекомендательных систем в области интеллектуального анализа данных. В работе представлены результаты исследования при формировании ансамбля алгоритмов для скоринговых систем с использованием слабоструктурированных данных, а предложенный ансамбль был протестирован на открытых данных портала UCI.

Введение

Одной из наиболее распространенных задач анализа данных является задача классификации. Задача классификации относится к разделу машинного обучения, который называется «обучение с учителем» (Supervised learning) [13]. Классификатором называется алгоритм, определяющий, какому из предопределенных классов принадлежит предъявляемый объект по вектору признаков. Подобный подход часто применяется в автоматизированных системах поддержки принятия решений таких, как рекомендательные системы, экспертные системы и т.д. [15].

Труды XVIII Международной конференции DAMDID/RCDL'2016 «Аналитика и управление данными в областях с интенсивным использованием данных», Ершово, 11-14 октября 2016

С ростом объемов данных, старые методы и способы обработки остаются в прошлом. Из-за обилия цифровой информации, поиск тематических статей или иных источников отнимает все больше времени и превращается в рутину. Зачастую, обработка данных должна выполняться в режиме онлайн, а требования к производительности и скорости работы являются довольно высокими.

Один из способов борьбы с такой проблемой являются рекомендательные и экспертные системы [12]. В своем ежегодном обзоре потенциала новых технологий и веяний, компания Gartner отмечает потенциал таких вещей, как умный советник, продвинутая аналитика с персональной доставкой информации, ответы на естественном языке, виртуальный персональный ассистент и прочее (см. рис.1[2]).

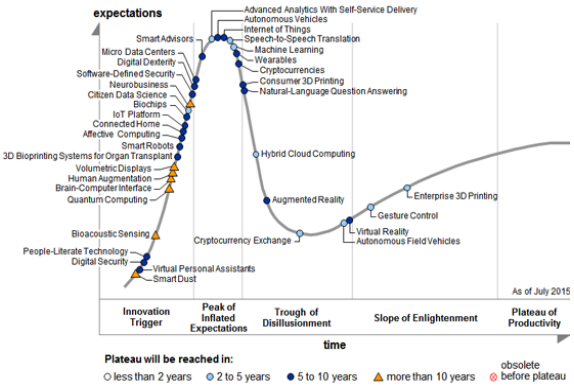

Рисунок 1 Цикл перспективных технологий

Использование описанных подходов нашло свое применение во многих областях, не исключая и финансовый сектор. Рост потребления благ человеком вынуждает в некоторых случаях использовать заемные средства. Любой заемщик для кредитной организации представляет определенный риск, а ее главной задачей является снижение такого риска. Появление систем оценки платежеспособности, основанное на численных статистических методах и дополненное средствами и инструментами машинного обучения, является одним из способов снижения риска для финансовой организации и роста ее доходов. Когда речь идет о больших суммах, рост точности прогнозного

значения даже на 1% может принести финансовой организации значительный доход.

Источниками данных для таких систем выступают статистические данные банков и иных кредитных организаций о выполнении клиентами своих обязательств. По каждому клиенту собирается и обрабатывается информация о его зарплате, имеющихся активах, образовании, кредитной истории, платежах по текущему кредиту и прочие данные.

Помимо описанных параметров, для оценки кредитоспособности потенциального клиента могут использоваться данные, не имеющие четкой структуры – слабоструктурированные данные. Такие данные могут быть представлены в различных форматах (как текстовом, так и графическом) и входят в общий перечень документов для оценки платежеспособности человека. К таким данным можно отнести различного рода рекомендательные письма, отзывы, справки и иные документы в свободной форме.

Помимо документов, которые предоставляет клиент, существуют и базы кредитных историй, которые доступны всем финансовым организациям для оценки заемщика. Ввиду различных стандартов кредитных организаций, некоторые поля могут заполняться в свободной форме: назначение кредита (как уже полученного, так и запросы), обеспечение, причина отказа и иные поля.

Целью данной статьи является представление авторского подхода к формированию ансамбля алгоритмов. В качестве прикладной области применения задачи классификации будут рассмотрено практическое использование ансамбля алгоритмов для формирования прогноза в области кредитного скоринга, где важную роль играют именно аналитические и экспертные системы принятия решений.

1 Алгоритмы классификации

Одним из подходов к решению задачи классификации является усиление простых классификаторов путём комбинирования примитивных слабых классификаторов в один сильный. Под силой классификаторов в данном случае подразумевается эффективность (качество) решения задачи классификации [11]. Основная идея использования ансамблей классификатор идентична той, когда при принятии важного решения человек пытается получить несколько различных мнений о своей проблеме, и на этой основе принимать решения.

Для оптимизации решения задач классификации и повышения точности работы алгоритмов выделяют несколько областей для исследования (см. рис.2).

Первым подходом для улучшения результата классификации является использование различных алгоритмов, отличных по своей природе и происхождению [14]. Существует значительное большое количество алгоритмов классификации, использование которых могут давать совершенно различные результаты.

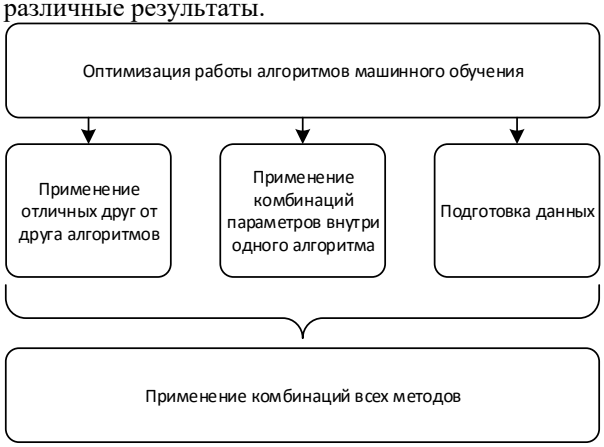

Рисунок 2 Методы оптимизации алгоритмов

Вторым подходом является оптимизация различных параметров известных алгоритмов, которые могут существенно влиять на результат. Например, в алгоритме RandomForest, такими параметрами являются [16]:

- количество деревьев входящих в алгоритм;
- минимальное число листов в дереве;
- минимальное расщепление в узле;
- минимальная и максимальная глубина;
- и другие.

В программной реализации алгоритма RandomForest число таких параметров значительно увеличивается за счет технических особенностей реализации и может доходить вплоть до 15.

Третьей подходом является работа с самими данными, которые подаются на вход алгоритма классификации. Подготовка данных – отдельная область, которая подразумевает обработку, очистку и приведение данных к машиночитаемому виду. Полученные данные могут быть изменены по какому-то признаку, могут быть добавлены новые значения, ранее не содержавшиеся в исходном наборе данных [8]. В некоторых случаях, данные могут быть удалены, т.к. не несут в себе полезной информации [4]. На основе исходных данных могут быть получены производные значения. Особенно это касается тех случаев, когда данные анонимны, т.е. представлены в виде цифровых значений или целиком в зашифрованном виде.

Таким образом, задача разработки оригинального алгоритма, который может являться ансамблем из нескольких известных алгоритмов, с различными весами и параметрами, и работающих со слабоструктурированными данными, является особо актуальной. При использовании данного подхода стоит обращать внимание на то, чтобы в итоговом ансамбле классификаторов были алгоритмы, имеющие разную природу происхождения [1]. Иначе, набор одинаковых алгоритмов будет выдавать схожие ответы и общее качество классификации значительно улучшить не получится.

Выделяют несколько способов для формирования ансамбля алгоритмов [6]:

- голосование большинством;
- веса пропорционально точности;
- использование условной вероятности;
- сероятностная формула Байеса;
- уменьшение дисперсии;
- независимость параметров друг от друга;
- взвешивание энтропии;
- плотностное взвешивание;
- и другие.

На сегодняшний день существует довольно много работ посвященных созданию скоринговых систем и различных комбинаций алгоритмов, позволяющих снизить риск невозврата кредита на основе данных о клиенте. Помимо широко распространённых и известных алгоритмов классификации, таких как Bagging и Boosting, создаются различные ансамбли алгоритмов классификации с предварительной кластеризацией [3], нечеткой логистической регрессией [9], а также использование нейросетей [7]. Представленные работы подробно описывают процесс, предшествующий принятию решения о выдаче кредита. В этом случае, используемые наборы данных в своей структуре содержат только количественные, категориальные и бинарные переменные.

Однако среди документов могут быть представлены и другие типы – текстовые данные. К ним могут относиться различного рода справки, анкеты, рекомендации и прочие неструктурированные текстовые данные. Текстовые данные имеют иную природу происхождения и требуют особого подхода к обработке и классификации.

Скоринговые системы направлены на работу с клиентом непосредственно до подписания договора, а затем работа скоринговой системы фактически заканчивается. В таком виде описанные системы никак не затрагивают процесс сопровождения клиентов и отслеживания рисков невозврата уже после получения кредита. Однако выделяют тип кредитного скоринга, который направлен на решение данной проблемы. Такой тип можно назвать «поведенческим», а основной задачей – регулярное отслеживание клиента в течение всего времени действия кредитного договора [5]. В качестве развития концепции о «поведенческой» составляющей клиента можно использовать не только финансовую информацию, но и оперировать ситуационными данными о клиенте, учитывать эмоциональное состояние клиента и прочее. Источником данных могут послужить социальные сети и иные открытые интернет источники.

Таким образом, работа с неструктурированными текстовыми данными может способствовать снижению потенциального риска выдачи кредита, а также учитывать в операционной деятельности кредитной организации и вероятность невозврата денежных средств в течение всего времени действия договора.

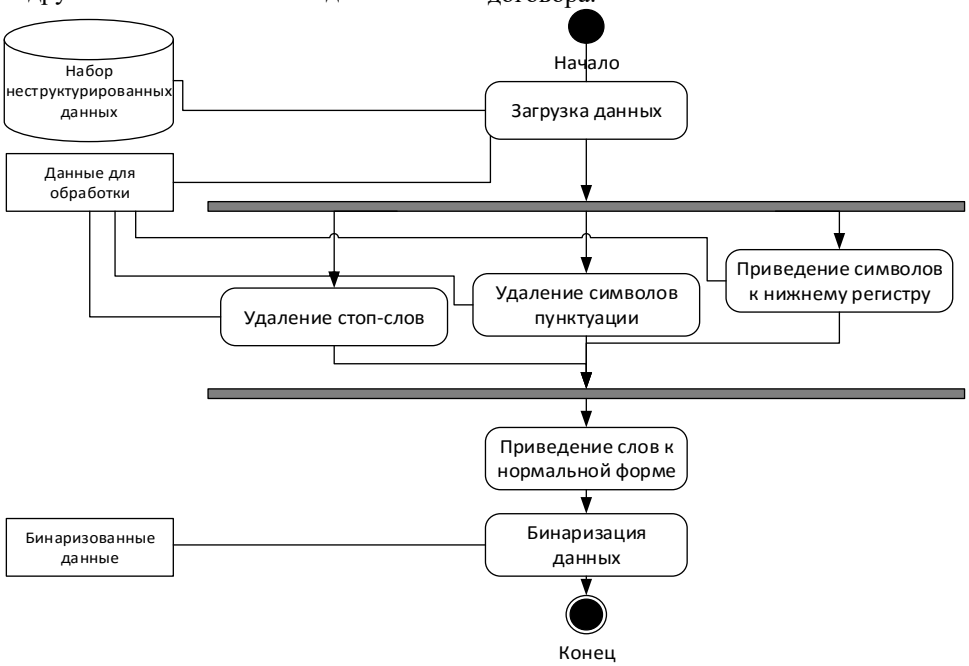

Рисунок 3 Предобработка текстовых данных

2 Ансамбль голосующих алгоритмов

В данной статье предлагается использовать следующие комбинации методов: использование условной вероятности и взвешивание энтропии.

В качестве основного коэффициента выступает энтропия, т.е. некая мера однородности результатов предсказания каждого алгоритма по отношению к правильному результату. Данная мера позволяет оценить качество работы конкретного классификатора для каждого класса.

Мера однородности рассчитывается по формуле (см. формулу 1):

$$H(x) = \sum_{i=1}^{n} p(i) \log_2 p(i) \qquad (1)$$

В качестве базовых алгоритмов могут рассматриваться любые простые алгоритмы, которые имеют различную природу происхождения. Это может быть комбинация алгоритмов «Случайный лес» (RandomForest - RF), «Наивный Байес» (Naïvebayes – NB), k-ближайших соседей (kNearestNeighbors – kNN) и других.

Также можно выделить классические ансамбли, которые зарекомендовали себя качеством и скоростью своей работы. К ним можно отнести классификаторы AdaBoost, Bagging и Boosting. Каждый из указанных ансамблей также можно встроить и использовать в других ансамблях.

Учитывая тот факт, что планируемый набор данных для классификации подразумевает работу со слабоструктурированными данными, то дополнительным шагом перед применением ансамбля алгоритмов является предобработка и выделение смысловой составляющей из слабоструктурированных данных.

На шаге предобработки загружаются неструктурированные данные, имеющие текстовый вид. Загруженные данные очищаются, нормализуются и приводятся к матричному виду. Алгоритм с предобработкой изображен на рисунке выше (см. рис.3).

Этап классификации состоит из трех шагов: обучение, тестирование, выбор результата (см. рис. 4).

На начальном этапе проводится обучение выбранных алгоритмов на основе обучающего набора данных. Выходными параметрами будут являться обученные модели по каждому классификатору.

Затем, на тестовой выборке проводится проверка обученных классификаторов и оценивается правильность полученных результатов. По каждому классификатору считается количество правильных и неправильных ответов. Для каждого алгоритма строится соответствующая матрица ошибок (см. матрицу 1):

$$\begin{pmatrix} n_{11} & n_{11} & \dots & n_{M1} \\ n_{21} & n_{11} & \dots & n_{M1} \\ \dots & \dots & \dots & \dots \\ n_{M1} & n_{M1} & n_{M1} & n_{MM} \end{pmatrix} \qquad (1)$$

Где n_{ij} – количество объектов, относящихся к i-ому классу, но классифицированных как класс j.

На основе представленной матрицы рассчитывается показатель энтропии для каждого алгоритма j по каждому классу i (см. матрицу 2):

$$\begin{pmatrix} H_{11} & H_{12} & \dots & H_{1N} \\ H_{21} & H_{21} & \dots & H_{2N} \\ \dots & \dots & \dots & \dots \\ H_{M1} & H_{M1} & \dots & H_{MN} \end{pmatrix} \qquad (2),$$

где H_{ij} – энтропия, M – число классов, N – число алгоритмов.

На следующем шаге выполняется расчет итогового класса на основе полученных значений. На тестовых данных считается вероятность по каждому алгоритму для каждого класса, затем полученные значения поочередно умножаются на соответствующую энтропию этого класса (при этом энтропия вычитается из единицы) (см. формулу 2).

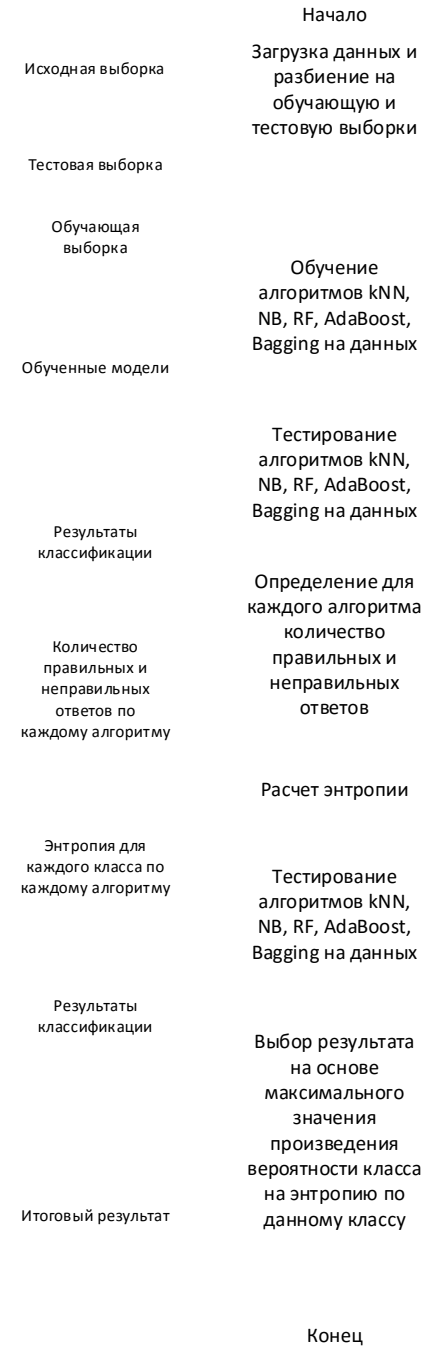

Рисунок 4 Алгоритм обучения

$$K = (1\text{-}p)*H \qquad (2),$$

где p – вероятность, H – энтропия, K – класс.

Затем выбирается класс с максимальным значением показателя качества K среди всех алгоритмов. Процедура проделывается для каждого

объекта выборки, до тех пор, пока не будет просмотрен весь список.

3 Тестирование алгоритма

В качестве предметной области для проверки результатов работы алгоритма был выбран набор данных, содержащих перечень кредитных платежей по клиенту в течение определенного периода времени (портал открытых данных UCI [10]). Представленный набор содержит базовую информацию о клиенте, получившем кредит: сумма, пол, образование, семейный статус и возраст. Помимо этого представленный набор данных содержит историю по платежам и суммам за полгода, что может позволить в дальнейшем сформировать некоторые паттерны для «поведенческого» типа скоринговых систем.

Задача: обучить классификатор и рассчитать точность его работы на тестовой выборке. Размер выборки составляет 30 тысяч различных записей. Количество классов в выборке равно двум: факт оплаты и факт неоплаты по кредитным обязательствам. Количество признаков равно 23. В качестве базовых алгоритмов были использованы: RandomForest, NaiveBayes, kNearestNeighbors, AdaBoost и Bagging.

После обучения базовых алгоритмов на 70% записей от исходного объема, и проведения тестирования на оставшихся 30% данных, была проведена серия экспериментов и были получены следующие результаты (см. таб. 1-4):

Таблица 1 Сравнение точности алгоритмов (первая итерация)

Алгоритм	Параметры	Точность
kNearestNeighbors	Кол-во соседей: 5	74,62%
RandomForest	Кол-во деревьев: 50	81,27%
NaiveBayes	Распределение: Бернулли	76,73%
AdaBoost	Кол-во классификаторов: 50	81,21%
Bagging	Кол-во классификаторов: 10	80,18%
Предлагаемый ансамбль		83,03%

Таблица 2 Сравнение точности алгоритмов (вторая итерация)

Алгоритм	Параметры	Точность
kNearestNeighbors	Кол-во соседей: 15	76,58%
RandomForest	Кол-во деревьев: 100	81,46%
NaiveBayes	Распределение: Бернулли	76,73%
AdaBoost	Кол-во классификаторов: 100	81,33%
Bagging	Кол-во классификаторов: 50	81,42%
Предлагаемый ансамбль		83,12%

Таблица 3 Сравнение точности алгоритмов (третья итерация)

Алгоритм	Параметры	Точность
kNearestNeighbors	Кол-во соседей: 40	77,47%
RandomForest	Кол-во деревьев: 200	81,48%
NaiveBayes	Распределение: Бернулли	76,73%
AdaBoost	Кол-во классификаторов: 200	81,30%
Bagging	Кол-во классификаторов: 75	81,22%
Предлагаемый ансамбль		83,25%

Таблица 4 Сравнение точности алгоритмов (четвертая итерация)

Алгоритм	Параметры	Точность
kNearestNeighbors	Кол-во соседей: 75	77,36%
RandomForest	Кол-во деревьев: 300	81,44%
NaiveBayes	Распределение: Бернулли	76,73%
AdaBoost	Кол-во классификаторов: 300	81,30%
Bagging	Кол-во классификаторов: 100	81,27%
Предлагаемый ансамбль		83,07%

Из таблицы видно, что предложенный алгоритм показывает сравнительно высокую точность, превосходящую точность метода случайного леса RF, а также алгоритмов AdaBoost и Bagging.Визуальное представление данных результатов представление можно увидеть на графике (см. рис.5).

В качестве направлений улучшения разработанного ансамбля планируются дальнейшие эксперименты с различными способами предобработки данных. Кроме того, планируется исследование применимости данного алгоритма на текстовых данных в области Web Mining и обработки сообщений из социальных сетей с целью извлечения

информации, способствующей оценки рисков потенциального заемщика.

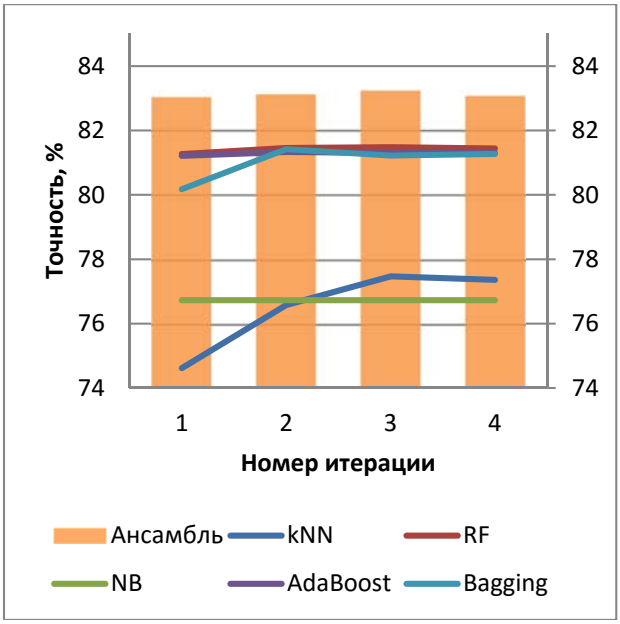

Рисунок 5 Сравнение точности предложенного ансамбля и классических подходов

4 Заключение

Решение задачи классификации востребовано во многих областях, связанных с обработкой больших объёмов данных, для поддержки процесса принятия решений. Существует множество алгоритмов классификации, эффективность работы которых ограничена объёмом и структурой данных, поэтому современный подход к данной проблеме заключается в конструировании ансамблей или комитетов более слабых алгоритмов. В данной работе предложен новый вариант ансамбля, использующий энтропийную меру в качестве меры однородности. Проведён эксперимент на открытых данных, который позволяет сделать заключение о перспективности предложенного метода классификации. Дальнейшие исследования по доработке разработанного ансамбля алгоритмов и его тестированию на слабоструктурированных данных поддержаны финансированием в рамках проекта № 57614X068 Федеральной Целевой Программы «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 годы».

Литература

[1] TG Dietterich Machine-learning research - Four current directions. AI MAGAZINE Том: 18 Выпуск: 4, 1997, с.: 97-136.

[2] Gartner's 2015 Hype Cycle for Emerging Technologies Identifies the Computing Innovations That Organizations Should Monitor. http://www.gartner.com/newsroom/id/3114217 (дата обращения: 22.05.2016)

[3] Hongshan Xiao, Zhi Xiao, Yu Wang. Ensemble classification based on supervised clustering for credit scoring. Appl. Soft Comput. 43: 73-86 (2016)

[4] John P. Cunningham, Yu Byron M Dimensionality reduction for large-scale neural recordings. NATURE NEUROSCIENCE. Том: 17, Выпуск: 11, 2014, с.: 1500-1509.

[5] Kenneth Kennedy, Brian Mac Namee, Sarah Jane Delany, M. O'Sullivan, N. Watson. A window of opportunity: Assessing behavioural scoring. Expert Syst. Appl. 40(4): 1372-1380 (2013)

[6] Lior Rokach. Ensemble-based classifiers. Springer Science+Business Media B.V., 2009.

[7] Petr Hájek, Vladimír Olej. Intuitionistic Fuzzy Neural Network: The Case of Credit Scoring Using Text Information. EANN 2015: 337-346

[8] Priti Gupta, Omdutt Sharma. – Feature selection: an overview. International Journal of Information Engineering and Technology (IMPACT: IJIET) ISSN(E): Applied; ISSN(P): Applied Vol. 1, Issue 1, Jul 2015, 1-12

[9] So Young Sohn, Dong Ha Kim, Jin Hee Yoon. Technology credit scoring model with fuzzy logistic regression. Appl. Soft Comput. 43: 150-158 (2016)

[10] UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients (дата обращения: 25.06.2016)

[11] Yeh, I. C., & Lien, C. H. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications, 36(2), 2009, p. 2473-2480.

[12] А.И. Гусева, В.С. Киреев, П.В. Бочкарёв, И.А. Кузнецов. Исследование алгоритмов многомерной классификации научных данных//Фундаментальные исследования. – 2015. – № 11(5). – С: 868-874.

[13] В.И. Донской Алгоритмические модели обучения классификации: обоснование, сравнение, выбор. Издательство «ДИАЙПИ», Симферополь, 2014

[14] М. П. Кривенко, В. Г. Васильев Методы классификации данных большой размерности. – М.: ИПИ РАН, 2013. 204 с.

[15] С.А. Филиппов, В.Н. Захаров, С.А. Ступников, Д.Ю. Ковалев. Организация больших объемов данных в рекомендательных системах поддержки жизнеобеспечения, входящих в состав глобальных платформ электронной коммерции. Институт проблем информатики ФИЦ ИУ РАН. Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), С. 119-124

[16] С.П. Чистяков Случайные леса: Обзор. Труды Карельского научного центра Российской академии наук, № 1, 2013, с. 117-136

Development of an ensemble of classification algorithms using the entropy quality measure for solving the problem of behavioral scoring

Igor A. Kuznetsov, Vasiliy S. Kireev

With increase of the amount of digital information the importance of a task of filtering and handling of such data increases in the world. For the purpose of identification of really required useful information for the user, the approaches based on the principles of classification of objects and rating of initial object to a group of the most similar to it are applied. Algorithms of machine learning act as a basis for classification, and classification itself is successfully applied in various data intensive areas, in particular, in the recommender systems. This article is devoted to the description of development of an ensemble of classification algorithms in case of creation of recommender systems in the field of data mining. In this article, the results of research for the case of forming ensemble of algorithms for scoring systems with use of semistructured data are presented, and the offered ensemble has been tested on open data of the UCI portal.