СРАВНЕНИЕ ЭФФЕКТИВНОСТИ МЕТОДОВ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В ЗАДАЧАХ МЕДИЦИНСКОЙ ДИАГНОСТИКИ[1]

Т.С. Карасева, студент, Д.Ю. Мамонтов, студент
Научный руководитель – Е.С. Семенкин, д.т.н., профессор
Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева, г. Красноярск
E-mail: tatyanakarasewa@yandex.ru

Очень часто перед врачом при диагностике заболеваний со схожими симптомами стоит проблема определения принадлежности данной болезни к тому или иному виду. Поэтому для решения большого числа медицинских задач применяются методы классификации. Ценность систем классификации в том, что они способны мгновенно анализировать и обобщать огромное количество прецедентов — возможность, недоступная специалисту-врачу.

В ходе исследовательской работы были протестированы методы классификации при решении трех медицинских задач. Для решения первой задачи были выбраны исходные данные, касающиеся диагностики патологии сердца. Следующая задача состоит в определении людей страдающих болезнью Паркинсона. Третья задача состоит в определении класса: здоровый человек, гипотиреоз, гипертиреоз [1].

Анализ данных осуществлялся с помощью системы Rapid Miner [2]. Были созданы модели относительно целевого атрибута для каждой задачи. В ходе работы был создан процесс, содержащий методы классификации: наивный байесовский классификатор (NB); метод k ближайших соседей (k-NN); деревья решений (DT); индукция правил (IR); логистическая регрессия (LR); машина опорных векторов (SVM); нейронная сеть (ANN); линейный дискриминантный анализ (LDA). С помощью оператора T-Test проведено сравнение используемых методов, чтобы увидеть, имеется ли между ними статистически значимое различие.

Наибольшей эффективностью при решении первой задачи обладает NB - 84%, хотя его точность классификации явно не может быть признана достаточной. К тому же его превосходство над методом SVM не является статистически значимым. Относительно второй задачи, можно сказать, что наибольшей эффективностью обладает ANN - 93%, однако его точность классификации не может быть признана наилучшей, так как он не имеет статистически значимого различия с SVM - 88% и RI - 88%. Исходя из полученных данных, можно утверждать, что наибольшей эффективностью при решении 3 задачи обладают методы: ANN - 99%, SVM - 99%, NB - 99%, LR - 96%, IR - 97%.

Далее был применен оператор ансамблирования Vote, использующий большинство голосов нескольких методов классификации, объединенных в ансамбль.

Для задачи по определению типа кардиалгии в ансамбль были включены SVM, NB, ANN, LR и IR, показавшие сопоставимую точность. Результат составил 91,19%. ANN, LDA, LR, NB, DT составили ансамбль для решения задачи по определению наличия болезни Паркинсона. Точность составила 90,18%. В ансамбль для решения задачи по определению типа состояния щитовидной железы были включены методы: NB, ANN, LR и IR. Результат равен 100%.

Последние мета-методы, которые были опробованы, это Баггинг (BG) и Бустинг (BT). Далее приведены результаты для всех поставленных задача. В скобках указаны алгоритмы, для которых применялись выбранные мета-методы. Задача №1: BG (NB) -

1 Работа выполнена в рамках и при финансовой поддержке проекта RFMEFI57414X0037.

86,75%, BG(SVM) - 88,07%; BT (NB) - 87,02%, BT(SVM) - 85,87%. Задача №2: BG (ANN) - 95,16%, BG(SVM) - 91,13%, BG(IR) - 90,93%; BT(ANN) - 94,10%, BT(SVM) - 89,11%, BT(IR) - 90,83%. Задача №3: BG (ANN) - 100%, BG(SVM) - 100%, BG(IR) - 100%; BT(ANN) - 100%, BT(SVM) - 100%, BT(IR) - 100%.

Ошибка при медицинской диагностике должна быть минимальна. Из этого следует, что методы интеллектуального анализа данных, реализованные в пакете Rapid Miner, далеко не всегда позволяют построить достаточно эффективные системы медицинской диагностики, т.е. необходимо их модифицировать и разрабатывать более мощные интеллектуальные технологии анализа данных. Одним из перспективных направлений является разработка технологий автоматизированного проектирования классификаторов на нечеткой логике [3], искусственных нейронных сетей [4], а также других методов анализа данных, с применением самонастраивающихся адаптивных алгоритмов оптимизации и моделирования [5] для выбора их эффективных структур и настройки параметров.

Список литературы:

1. Machine Learning Repository [Электронный ресурс]. URL:http://archive.ics.uci.edu/ml/ datasets.html (дата обращения: 4.12.2014).
2. RapidMiner [Электронный ресурс]. URL: https://rapidminer.com/ (дата обращения: 18.12.2014).
3. Semenkin E., Stanovov V. Fuzzy Rule Bases Automated Design with Self-configuring Evolutionary Algorithm // Informatics in Control, Automation and Robotics (ICINCO), 11th International Conference on. INSTICC, 2014. – Vol. 1. – P. 318-323.
4. Akhmedova Sh., Semenkin E. Co-operation of Biology Related Algorithms meta-heuristic in ANN-based classifiers design // Proceedings of the IEEE Congress on Evolutionary Computation 2014. - P. 867-872.
5. Semenkin E.S., Semenkina M.E. Self-configuring Genetic Algorithm with Modified Uniform Crossover Operator // Advances in Swarm Intelligence. Lecture Notes in Computer Science 7331. – Springer-Verlag, Berlin Heidelberg, 2012. – P. 414-421.