ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ ДЛЯ ПРОГНОЗИРОВАНИЯ УСПЕШНОСТИ УЧЕБНОЙ ДЕЯТЕЛЬНОСТИ АБИТУРИЕНТОВ

Ш. А. Ахмедова, С. Р. Вишневецкая, А. А. Коромыслова

Сибирский государственный аэрокосмический университет имени академика М. Ф. Решетнева
Российская Федерация, 660014, г. Красноярск, просп. им. газ. «Красноярский рабочий», 31
E-mail: shahnaz@inbox.ru, akoromyslova@mail.ru, vishni@ngs.ru

Представлено решение задачи прогнозирования успешности обучения абитуриентов в случае их зачисления в университет. Для этого были применены различные методы классификации и прогнозирования, а именно: машины опорных векторов, нейронные сети и нечеткие системы, настроенные генетическими и стайными «бионическими» алгоритмами. Для решения указанной задачи были предварительно собраны и преобразованы данные, полученные за время работы приемной комиссии. Кроме того, изначально была решена задача классификации, а именно, поступит или нет выпускник в университет. В итоге было установлено, какие именно сведения, предоставленные абитуриентами, достаточны для определения успешности прохождения первой экзаменационной сессии теми из них, кто поступит в вуз. Результаты решения задачи, полученные методом опорных векторов, оказались наилучшими.

Ключевые слова: успеваемость студентов первого курса, метод опорных векторов, нейронные сети, нечеткие системы, генетические алгоритмы, стайные «бионические» алгоритмы.

Vestnik SibGAU
2014, No. 3(55), P. 16–20

DATA MINING TOOLS FOR PROSPECTIVE STUDENTS' SUCCESS RATE PREDICTION

Sh. A. Akhmedova, S. R. Vishnevskaya, A. A. Koromyslova

Siberian State Aerospace University named after academician M. F. Reshetnev
31, Krasnoyarsky Rabochy Av., Krasnoyarsk, 660014, Russian Federation
E-mail: shahnaz@inbox.ru, akoromyslova@mail.ru, vishni@ngs.ru

The aim of that study was prospective students' success rate prediction in case if they are enrolled at university. For this purpose different classification and prediction methods were used namely support vector machines (SVM), neural networks and fuzzy systems which were obtained by using genetic and bionic algorithms and their modifications. Firstly, meta-heuristic called Co-Operation of Biology Related Algorithms (COBRA), that has earlier demonstrated its usefulness on CEC'2013 real-valued optimization competition benchmark and its modification for solving constrained optimization problems, was applied to SVM-based classifiers design. Then genetic algorithm was used for neural networks and fuzzy logic systems automatic generation. Various benchmark classification problems were solved with those approaches. It was established that support vector machines, neural networks and fuzzy logic systems developed in that way outperform many alternative methods on mentioned benchmark classification problems. So the workability and usefulness of proposed classification or prediction algorithms were confirmed. After that for solving of prospective students' success rate prediction problem, in case if they are enrolled at university, information about them obtained during the operational period of Admission Committee was gathered and preprocessed. Eventually it was established which kind of information about a prospective student is enough for determining whether he or she will or will not pass end of first semester exams in case of enrollment. Moreover, it should be noted that support vector machines generated by collective optimization method COBRA have shown the best results.

Keywords: fresh students' success rate, support vector machines, neural networks, fuzzy systems, genetic algorithms, bionic algorithms.

Введение. Подготовка квалифицированных специалистов – одна из главных задач любого образовательного учреждения. Подходом, позволяющим хотя бы приблизительно предсказать возможность успешного

поступления и сдачи, к примеру, первых экзаменов (первой экзаменационной сессии), является прогнозирование успешности учебной деятельности студента по информации, доступной на момент его поступления

в вуз. Данная информация сосредоточена в документах приемной комиссии и представляет собой набор разнообразных сведений, выраженных в различных шкалах, чаще всего номинальных (пол, номер школы, город и т. п.) и порядковых (выбранные приоритеты), реже – в дискретных (возраст, результаты сдачи ЕГЭ) и совсем редко – в абсолютных. Результаты обучения студентов, в свою очередь, сосредоточены в документах деканатов и представляют собой сведения в дискретных или порядковых шкалах (оценки, результаты аттестации), а иногда – в номинальных (благодарности, выговоры и т. п.). В данной работе были использованы лишь сведения о поступлении абитуриентов и факте продолжения обучения студента в следующем семестре или его отчислении по результатам сдачи экзаменов.

На успеваемость студента могут влиять многие факторы. Можно выделить следующую группу характеристик для оценивания «перспективности» абитуриента: пол абитуриента, его возраст, адрес проживания, номер школы, выпустившей абитуриента, результаты сдачи ЕГЭ, олимпиад, вступительных экзаменов в вузе (если нет соответствующих результатов по ЕГЭ), выбранные приоритеты при поступлении в вуз. Иными словами, учитывая всю информацию, предоставляемую студентом при поступлении, возможно спрогнозировать, какой в дальнейшем будет его успеваемость, удастся или нет ему пройти хотя бы часть обучения.

Составить единую формальную модель влияния перечисленных факторов на успеваемость будущих студентов непросто: нужно рассматривать как количественные показатели, так и качественные, поэтому можно сказать, что исходные данные являются сложно формализуемыми. Необходимо собрать сведения для формирования базы данных, которая бы использовалась для генерирования интеллектуальной информационной технологии прогнозирования. Таким образом, необходимо построить модель, описывающую связь между группой характеристик абитуриента и его дальнейшей успеваемостью в вузе в случае его поступления, основываясь на сформированной и обновляемой базе данных.

Для этой цели наиболее подходящими инструментами являются искусственные нейронные сети (Artificial Neural Networks, ANN) [1], метод опорных векторов (Support Vector Machine, SVM) [2], а также нечеткие системы [3]. В данной работе предлагается генерирование интеллектуальных информационных технологий прогнозирования и классификации на основе перечисленных методов, настроенных коллективным бионическим алгоритмом стайного интеллекта (Co-Operation of Biology Related Algorithms, COBRA) [4] или генетическим алгоритмом [5]. Главная идея метода COBRA заключается в параллельной работе пяти известных алгоритмов, а именно, метода опорных векторов (Particle Swarm Optimization, PSO) [6], алгоритма поиска стаей волков (Wolf Pack Search, WPS) [7], алгоритма летучих мышей (Firefly Algorithm, FFA) [8], алгоритма поиска кукушек (Cuckoo

Search Algorithm, CSA) [9] и алгоритма летучих мышей (Bat Algorithm, BA) [10]. Итак, упомянутый алгоритм COBRA был предложен для автоматизации проектирования SVM-«машин», так как по сравнению с обычно применяемыми роевыми алгоритмами он обладает более высокой эффективностью при решении оптимизационных задач с вещественными переменными [11]. Генетический же алгоритм был предложен для автоматизации проектирования нейросетей и нечетких систем. Помимо прочего, также был автоматизирован процесс поиска наиболее информативных данных (входов) при решении подобного рода задач. Подобные алгоритмы генерирования интеллектуальных информационных технологий были рассмотрены ранее в работах [12; 13].

Таким образом, сначала необходимо собрать сведения для формирования базы данных о поступавших абитуриентах. В этой базе данных должны быть их характеристики, а также сведения об их успеваемости во время обучения в вузе (было ли отчисление или нет). Кроме того, после сбора данных должны быть решены две задачи классификации:

- характеристики абитуриента – входные данные, а сведения о том, поступил он или нет – выходные (т. е. всего 2 класса);

- характеристики абитуриента – входные данные, показатели успеваемости – выходные параметры для будущей модели, т. е. определены 3 класса: поступивший абитуриент и сдавший первую экзаменационную сессию, поступивший абитуриент и отчисленный после первой сессии, непоступивший абитуриент.

Собранные данные должны быть предварительно обработаны таким образом, чтобы было возможно их дальнейшее использование при решении описанных задач предлагаемыми методами, а именно, приведены к численному виду и нормированы.

Метод опорных векторов. Метод опорных векторов [2] – алгоритм типа «обучение с учителем», изначально разработанный для решения задач бинарной категоризации. Идея алгоритма заключается в том, что каждый объект из набора данных представляется в виде вектора в пространстве, далее полученные векторы необходимо разделить некоторой гиперплоскостью таким образом, чтобы объекты разных классов были как можно дальше друг от друга. То есть сначала алгоритм обучается на некотором наборе данных, а затем категории новых объектов определяются проверкой, по какую сторону построенной гиперплоскости оказались соответствующие этим объектам векторы в пространстве.

Пусть дан набор данных:

$$X^l = \{(x_1, y_1), \dots, (x_l, y_l)\}, x_i \in R^m, y_i \in \{-1; 1\},$$

где l – это количество объектов в обучающей выборке с m вещественными атрибутами (таким образом, каждый объект в выборке можно представить как вектор размерности m); y_i – класс i -го объекта из выборки.

Как уже отмечалось, в самом общем виде метод опорных векторов заключается в том, что необходимо построить гиперплоскость

$$\langle w, x \rangle + b = 0,$$

где $w \in R^m, b \in R, \langle _, _ \rangle$ – скалярное произведение, которая разделяла бы объекты класса -1 от объектов класса 1 в пространстве. Тогда классификация нового объекта осуществляется по следующему правилу:

$$f(x) = \begin{cases} 1, & \langle w, x \rangle + b \geq 1, \\ -1, & \langle w, x \rangle + b \leq -1. \end{cases}$$

Помимо прочего, в методе опорных векторов предполагается, что расстояние между построенной гиперплоскостью и ближайшими к ней классифицируемыми объектами должно быть максимально возможным. Гиперплоскость с упомянутым выше свойством может быть получена при решении следующей задачи условной оптимизации:

$$\|w\|^2 \rightarrow \min,$$

$$y_i(\langle w, x_i \rangle + b) \geq 1, i = \overline{1, l}.$$

Таким образом, для решения задачи классификации методом опорных векторов необходимо определить вектор w и параметр смещения b , иначе говоря, решить задачу условной оптимизации для построения разделяющей гиперплоскости.

Но как было уже упомянуто, также определялись наиболее информативные входы для решаемой задачи. С этой целью для каждого вектора входных параметров генерировалась бинарная строка, количество бит в строке было равным количеству входов, так что некоторый бит в строке был равен единице, если соответствующий ему параметр стоило учитывать (он был «информативен»), и нулю – в противном случае. Следовательно, помимо построения оптимальной гиперплоскости нужно было решить задачу безусловной оптимизации с бинарными переменными. В итоге генерировалась популяция бинарных строк (различные способы учета входных данных), для каждой строки решалась задача условной оптимизации и строилась гиперплоскость, далее выбиралась строка с лучшим показателем эффективности. После описанного этапа уже выполнялась работа бинарного алгоритма.

В данной работе для настройки поиска информативных входов была применена модификация коллективного бионического алгоритма для решения задач с бинарными переменными, описанная в [14] (COBRA-b), а его модификация для решения задач условной оптимизации (COBRA-c) применялась для нахождения вектора w и параметра смещения b [15]. Итак, COBRA-b использовалась для определения информативных входов: каждый индивид описывал набор параметров. Далее для каждого такого индивида с помощью модификации алгоритма COBRA для решения задач условной оптимизации осуществлялся поиск

вектора w и параметра смещения b . В конце выбирался индивид с наилучшим показателем эффективности классификации (т. е. с минимальной ошибкой классификации).

Нейронные сети и системы на нечеткой логике.

Для данной работы были реализованы нейронные сети с настраиваемыми структурой и весовыми коэффициентами, а также выбором наиболее информативных входов. Настройка осуществлялась генетическими алгоритмами (ГА). Изначально ГА был использован для настройки весовых коэффициентов полностью связанных нейронных сетей прямой проводимости (GA-ANNW) с произвольным числом скрытых слоев и нейронов на них. В качестве активационной функции для нейронов на выходном и скрытых слоях применялась сигмоида. Весовые коэффициенты записывались в хромосому последовательно в виде двоичного кода. Количество используемых бит зависело от точности настройки и разброса возможных значений весов.

Далее решалась задача выбора структуры нейронной сети, а именно, необходимо было определить число слоев и количество нейронов на них, а также установить вид активационной функции каждого нейрона. Оптимальную структуру могут определить эксперты, но это потребует больших временных затрат и материальных ресурсов, поэтому предлагается вновь использовать генетический алгоритм для автоматического проектирования структуры нейросети (GA-ANNS).

Таким образом, скрытые слои кодировались последовательно. Каждый нейрон был закодирован в четырех битах. Для каждого нейрона сначала случайным образом, с определенной вероятностью, равной $1/3$, решалось, будет ли он использоваться в данной сети. Если нейрона в сети не было, то на его место в хромосому записывались нули. В противном случае для него случайным образом выбиралась одна из пятнадцати функций активации [16], номер которой записывался в двоичном коде. Далее для каждой выбранной структуры нейронной сети запускался генетический алгоритм GA-ANNW для выбора весовых коэффициентов.

Так как эффективность применения генетического алгоритма для ИНС зависела от размерности задачи, было целесообразно не учитывать малоинформативные признаки. Особенность генетического алгоритма с выбором наиболее информативных признаков при автоматической настройке нейронной сети (GA-ANNinput) заключается в использовании ГА бинарной хромосомы, в которую нейронная сеть кодируется последовательно: сначала записываются все входные переменные (по одному биту на признак), а затем нейроны со скрытых слоев, кодируемые аналогично GA-ANNS. При инициализации признак считался неинформативным с вероятностью $0,5$.

Генетический алгоритм, настраивающий нейронные сети, решает задачу на приемлемом уровне, однако является «черным ящиком». Поэтому предлагается также использовать системы на нечеткой логике. При

работе с системами на нечеткой логике возникают две проблемы: определение способа кодирования лингвистических переменных и выбор оптимальной базы правил. Лучший способ решения этих проблем – работа с экспертом. Однако это требует материальных и временных затрат, а иногда и вообще не представляется возможным. Поэтому предлагается использовать генетический алгоритм для автоматической настройки систем, основанных на нечёткой логике (GA-FL).

Для кодирования лингвистических переменных генерировалась бинарная строка, к которой далее применялись операторы генетического алгоритма. Все нечеткие переменные описываются набором (N, X, A) , где N – это название переменной, X – универсальное множество (область рассуждений), A – нечеткое множество на X .

В данной работе одновременно использовались три вида функции принадлежности: треугольная, трапециевидная и гауссова. Для формирования базы правил также использовалась бинарная хромосома. Количество битов на одно правило зависело от размерности задачи.

Было установлено, что тонность GA-FL заметно падает при росте размерности. Чтобы оставить точность на приемлемом уровне, используя меньше ресурсов, предлагается производить выбор наиболее информативных признаков. Генетический алгоритм для автоматической настройки систем на нечёткой логике с выбором наиболее информативных признаков (GA-FLinput) основан на том же принципе выбора информативных признаков, что и GA-ANNinput.

Прогнозирование успешности учебной деятельности абитуриентов. Описанные ранее алгоритмы были применены для решения двух задач классификации. Первая задача заключалась в определении, поступит или нет абитуриент, соответственно, задача с двумя классами – поступившие и не поступившие абитуриенты. Вторая задача состояла в прогнозировании возможности отчисления или прохождения первой экзаменационной сессии поступившими абитуриентами. Всего для второй задачи были определены три класса: поступившие выпускники школ, отчисленные после первой экзаменационной сессии; поступившие выпускники школ, сдавшие первую сессию; не поступившие абитуриенты.

Для обеих задач использовался один и тот же набор данных, состоявший из сведений об 742 абитуриентах. Каждый поступающий описывался следующими характеристиками: возраст, адрес, пол, баллы за ЕГЭ, сдавали ли экзамен в университете и т. д. Итак, входных параметров было 11, причем первый параметр был идентификатором класса. Все данные были изначально нормированы.

И для первой, и для второй задачи алгоритмы запускались по 20 раз. В таблице приведены полученные результаты, усредненные по количеству запусков, причем в таблице для каждого алгоритма указана доля правильно классифицированных объектов, т. е. абитуриентов.

Результаты работы алгоритмов

Задача	GA-ANNinput	GA-FLinput	SVM+COBRA
2 класса	0,973	0,962	0,991
3 класса	0,931	0,915	0,969

По полученным результатам можно сделать вывод, что на данных задачах метод опорных векторов в среднем работает лучше нейронных сетей и систем на нечеткой логике, причем нейронные сети оказались эффективней метода GA-FLinput. Также стоит отметить, что все время работы всех алгоритмов (время, необходимое для выполнения одного прогона) было примерно одинаковым (около 3–5 минут).

Одной из целей данной работы было определение наиболее информативных входов для этих задач, т. е. определение самых значимых факторов для поступления и сдачи первой экзаменационной сессии. Так, например, наилучший результат (98,7 % правильно классифицированных данных) для задачи с тремя классами был достигнут методом SVM+COBRA при учете следующих входных параметров: вид образовательного учреждения, оконченного абитуриентом, возраст абитуриента, баллы за ЕГЭ по информатике. Для первой же задачи наилучший результат достигался при учете тех же входных параметров, только вместо баллов за ЕГЭ по информатике включались баллы за ЕГЭ по математике. Обращает на себя внимание тот факт, что существенным фактором оказался возраст, хотя, казалось бы, у абитуриентов он практически одинаков. Объяснением этого может быть тот факт, что более старшие абитуриенты, окончившие школу не в текущем году, а ранее, менее успешны при поступлении и дальнейшем обучении. Очевидно, что среди таких поступающих значительная доля не поступивших в прошлые годы или поступивших и отчисленных после первых сессий, т. е. менее пригодных для обучения в вузе.

Заключение. Таким образом, были решены две задачи классификации успешности поступления и учебной деятельности абитуриентов тремя различными способами: методом опорных векторов, нейронными сетями и системами на нечеткой логике, настроенными бионическими, а также генетическими алгоритмами. В итоге было установлено, что для решения обеих задач необходимо учитывать следующие факторы: вид образовательного учреждения, оконченного абитуриентом, и его возраст. Кроме того, в зависимости от решаемой задачи следует также использовать сведения о баллах за ЕГЭ либо по математике (для задачи с 2 классами), либо по информатике (соответственно, для второй задачи с 3 классами).

Библиографические ссылки

1. Bishop C. M. Theoretical foundation of neural networks // Proceedings of Physics Computing 1996. P. 500–507.
2. Вапник В., Червоненкис А. Теория распознавания образов. М. : Наука, 1974. 415 с.

3. Круглов В. В., Дли М. И., Голунов Р. Ю. Нечеткая логика и искусственные нейронные сети. М.: Физматлит, 2001. 224 с.
4. Akhmedova Sh., Shabalov A. Development and Investigation of Bio-logically Inspired Algorithms Cooperation Metaheuristic // Proceedings of the Genetic and Evolutionary Computation Conference Companion (GECCO'13) 2013. P. 1417–1418.
5. Гуменникова А. В. [и др.]. Об эволюционных алгоритмах решения сложных задач оптимизации // Вестник СибГАУ. 2003. № 4. С. 14–23.
6. Kennedy J., Eberhart R. Particle Swarm Optimization // Proceedings of IEEE Intern. Conf. on Neural Networks, 1995. Vol. IV. P. 1942–1948.
7. Yang Ch., Tu X., Chen J. Algorithm of Marriage in Honey Bees Optimization Based on the Wolf Pack Search // Proceedings of the Intern. Conf. on Intelligent Pervasive Computing. 2007. P. 462–467.
8. Yang X. S. Firefly algorithms for multimodal optimization // Proceedings of the 5th Symposium on Stochastic Algorithms, Foundations and Applications. 2009. P. 169–178.
9. Yang X. S., Deb S. Cuckoo Search via Levy flights // Proceedings of the World Congress on Nature & Biologically Inspired Computing. 2010. P. 210–214.
10. Yang X. S. A new metaheuristic bat-inspired algorithm // Nature Inspired Cooperative Strategies for Optimization, Studies in Computational Intelligence. 2010. Vol. 284. P. 65–74.
11. Akhmedova Sh., Semenkin E. Co-Operation of Biology Related Algorithms // Proceedings of the IEEE Congress on Evolutionary Computation. 2013. P. 2207–2214.
12. Семенкина М. Е. Самоадаптивные эволюционные алгоритмы проектирования информационных технологий интеллектуального анализа данных // Искусственный интеллект и принятие решений. 2013. № 1. С. 13–23.
13. Семенкин Е. С., Шабалов А. А. Система автоматизированного проектирования коллективов интеллектуальных информационных технологий для задач анализа данных // Программные продукты и системы. 2012. № 4 (100). С. 70–73.
14. Ахмедова Ш. А., Семенкин Е. С. Новый коллективный метод оптимизации на основе кооперации бионических алгоритмов // Вестник СибГАУ. 2013. № 4 (50). С. 92–99.
15. Akhmedova Sh., Semenkin E., Sergienko R. Automatically Generated Classifiers for Opinion Mining with Different Term Weighting Schemes // Proceedings of the 11th Intern. Conf. on Informatics in Control, Automation and Robotics (ICINCO'2014). 2014. Vol. 2. P. 845–850.
16. Electronic textbook StatSoft [Электронный ресурс]. URL: <http://www.fmi.uni-sofia.bg/fmi/statist/education/textbook/eng/glosa.html>.
2. Vapnik V., Chervonenkis A. *Teoriya raspoznavaniya obrazov* [The Theory of Pattern Recognition]. Moscow, Nauka Publ., 1974, 415 p.
3. Kruglov V. V., Dlee M. I., Golunov R. Y. *Nechetkaya logika I iskusstvennye neironnye seti* [Fuzzy logic and artificial neural networks]. Moscow, Phismatlit Publ., 2001, 224 p.
4. Akhmedova Sh., Shabalov A. Development and Investigation of Bio-logically Inspired Algorithms Cooperation Metaheuristic. Proceedings of the *Genetic and Evolutionary Computation Conference Companion (GECCO'13)*. 2013, p. 1417–1418.
5. Gumennikova A. V., Emel'janova M. N., Semenkin E. S., Sopov E. A. [About evolutionary algorithms for solving hard optimization problems]. *Vestnik SibGAU*. 2003, no. 4, p. 14–23. (In Russ.)
6. Kennedy J., Eberhart R. Particle Swarm Optimization. Proceedings of *IEEE International Conference on Neural Networks*. 1995, vol. IV, p. 1942–1948.
7. Yang Ch., Tu X., Chen J. Algorithm of Marriage in Honey Bees Optimization Based on the Wolf Pack Search. Proceedings of the *International Conference on Intelligent Pervasive Computing*. 2007, p. 462–467.
8. Yang X. S. Firefly algorithms for multimodal optimization. Proceedings of the *5th Symposium on Stochastic Algorithms, Foundations and Applications*. 2009, p. 169–178.
9. Yang X. S., Deb S. Cuckoo Search via Levy flights. Proceedings of the *World Congress on Nature & Biologically Inspired Computing*. 2010, p. 210–214.
10. Yang X. S. A new metaheuristic bat-inspired algorithm. *Nature Inspired Cooperative Strategies for Optimization, Studies in Computational Intelligence*. 2010, vol. 284, p. 65–74.
11. Akhmedova Sh., Semenkin E. Co-Operation of Biology Related Algorithms. Proceedings of the *IEEE Congress on Evolutionary Computation (CEC'2013)*. 2013, p. 2207–2214.
12. Semenkina M. E. [Effectiveness investigation of adaptive evolutionary algorithms for data mining information technology design]. *Iskusstvennyj intellekt i prinjatie reshenij*. 2013, no. 1, p. 13–23. (In Russ.)
13. Semenkin E. S., Shabalov A. A. [The system of automatic design of intelligent information technologies ensembles for data mining problems]. *Programmnye produkty i sistemy*. 2012, no. 4 (100), p. 70–73. (In Russ.)
14. Akhmedova Sh. A., Semenkin E. S. [New optimization metaheuristic based on co-operation of biology related algorithms]. *Vestnik SibGAU*. 2013, no. 4 (50), p. 92–99. (In Russ.)
15. Akhmedova Sh., Semenkin E., Sergienko R. Automatically Generated Classifiers for Opinion Mining with Different Term Weighting Schemes. Proceedings of the *11th International Conference on Informatics in Control, Automation and Robotics (ICINCO'2014)*. 2014, vol. 2, p. 845–850.
16. Electronic textbook StatSoft. Available at: <http://www.fmi.uni-sofia.bg/fmi/statist/education/textbook/eng/glosa.html>.

References

1. Bishop C. M. Theoretical foundation of neural networks. Proceedings of *Physics Computing*. 1996, p. 500–507.

© Ахмедова Ш. А., Вишневецкая С. П., Коромыслова А. А., 2014