УДК 004.93

Отбор информативных признаков на основе квантовых вычислений

А. В. Комендант[1], С. А. Субботин[1]

Аннотация — **The method of feature selection based on quantum computing is offered. It can improve the quality and speed of diagnostic model synthesis.**

Ключевые слова — **feature selection, quantum computing, technical diagnosis.**

I. Введение

Для построения моделей зависимостей по набору наблюдений на основе индуктивных методов обучения необходимо обладать набором информативных признаков. При этом для упрощения моделей, а также ускорения процесса их построения целесообразно выделить минимальный набор признаков, обладающих совместно наибольшей информативностью.

Целью работы являлось создание метода отбора информативных признаков на основе квантовых вычислений для повышения точности модели и скорости ее построения.

II. Метод отбора информативных признаков на основе квантовых вычислений

Пусть мы имеем набор обучающих данных $\{<x^s,y^s>\}$, где $x^s=\{x^s_j\}$, $j=\overline{1,N}$, x^s_j – значение j-го входного, а y^s – выходного признака, сопоставленное s-му экземпляру выборки x^s. Задача отбора информативных признаков будет заключаться в нахождении такой комбинации признаков, для которой модель будет обладать наибольшей точностью. Отбор признаков предлагается осуществлять на основе стохастического поиска [1] и квантовых вычислений [2].

Будем считать, что вся популяция решений P во времени состоит из дискретных поколений (генераций, эпох): $P=\{P_t, t=\overline{1,T}\}$, где P_t – t-ое поколение особей; T – количество поколений. Каждое поколение состоит из N особей: $P_t=\{H_j\}, j=\overline{1,N}$, где H_j – j-ая особь популяции в t-ом поколении. Каждая хромосома (особь, точка в пространстве поиска) оценивается мерой приспособленности в соответствии с тем, насколько хорошим является соответствующее ей решение задачи. Приспособленность определяется как значение целевой функции (фитнесс-функции) $f(H_j)$ для каждой из хромосом. Каждая хромосома состоит из L генов: $H_j=\{H_{ij}\}, i=\overline{1,L}$, где H_{ij} – i-ый ген j-ой хромосомы.

В данном методе хромосомы состоят из генов, каждый из которых представляет собой кубит (квантовый бит). Каждый кубит имеет два выделенных состояния $|0\rangle$ и $|1\rangle$ (если считать кубиты спинами, то это состояния «спин вверх» и «спин вниз»). Указание выделенных состояний для каждого кубита системы задает не все возможные состояния системы, а только базисные. Возможны также любые линейные комбинации базисных состояний с комплексными коэффициентами.

Состояние кубита может быть представлено в следующем виде: $|\psi\rangle=\alpha|0\rangle+\beta|1\rangle$, где $|0\rangle$ и $|1\rangle$ — выделенные состояния кубита; α и β — комплексные числа, которые определяют вероятностные амплитуды соответствующих состояний. При любом измерении состояния кубита он случайно переходит в одно из своих выделенных состояний. Вероятности наблюдения кубита в его выделенных состояниях $|0\rangle$ и $|1\rangle$ соответственно равны $|\alpha|^2$ и $|\beta|^2$. В данном методе вероятностная амплитуда кубита определяется парой вещественных чисел α и β как $[\alpha\ \beta]^T$, причем на вероятностные амплитуды накладывается условие нормирования: $|\alpha|^2+|\beta|^2=1$, выражающее тот факт, что суммарная вероятность должна быть равна единице.

Фаза кубита определяется углом $\xi = \mathrm{arctg}(\beta / \alpha)$. Квадрант, в котором находится фаза кубита, определяется знаком произведения d вероятностных амплитуд кубита: $d = \alpha\cdot\beta$. Если $d > 0$, то ξ лежит в первом или в третьем квадранте, если $d < 0$, то ξ лежит во втором или в четвертом квадранте.

Для ускорения процесса поиска предлагается начальную точку поиска выбирать не случайным образом, а учитывать при ее выборе оценки индивидуальной информативности признаков I_j [1] при задании начальных значений α и β. С другой стороны для обеспечения глобального характера поиска и разнообразия решений при задании начальных значений этих параметров целесообразно сохранить случайную составляющую. Таким образом, начальные значения α и β предлагается выбирать случайным образом, но вероятностно зависящим от оценок индивидуальной информативности признаков. Также предлагается учесть оценки индивидуальной информативности признаков при выборе коэффициента k, определяющего скорость сходимости метода.

Работа метода может быть представлена следующей последовательностью этапов.

Этап 1. Установить счетчик итераций (времени): $t = 0$. Выполнить инициализацию начальной популяции особей P_t. Задать начальные значения α и β с учётом оценок индивидуальной информативности признаков I_j: $\alpha_{ij}=m\cdot I_j+(1-m)\mathrm{rand}, \beta_{ij}=\pm\sqrt{1-|\alpha_{ij}|^2}$, где I_j – оценка индивидуальной информативности j-го признака; rand — случайное число ($0\le \mathrm{rand}\le 1$); m – коэффициент, определяющий влияние оценок I_j ($0\le m\le 1$): при $m = 0$ оценки I_j не учитываются, при $m = 1$ начальная квантовая популяция P_0 будет состоять из одинаковых особей. Но даже при $m = 1$ и/или $\mathrm{rand}\equiv 0$ будет обеспечен глобальный характер поиска и многообразие решений вследствие того, что измерение

[1] Запорожский национальный технический университет, ул. Жуковского, 64, Запорожье, 69063, УКРАИНА, E-mail: subbotin@zntu.edu.ua

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua

состояния кубитов на следующем этапе является вероятностным процессом, результат которого будет содержать случайную составляющую.

Этап 2. В соответствии с вероятностными амплитудами α и β каждого гена каждой особи, сформировать популяцию решений P_t' путем измерения состояний кубитов-генов квантовой популяции P_t.

Этап 3. Для каждого решения построить модель по соответствующему набору признаков. Оценить особи текущей популяции P_t' путем вычисления для них значений фитнесс-функции $f(H_j), j = \overline{1,N}.$

Этап 4. Из текущей популяции выбрать лучшее решение H_b. Если лучшее на текущей итерации решение оказалось лучше, чем лучшее за всё время работы метода решение H_{best}, то принять $H_{best} = H_b$.

Этап 5. Проверить условия окончания поиска. В качестве таких критериев могут быть использованы: достижение максимально допустимого времени или числа итераций работы метода, либо достижение приемлемого значения фитнесс-функции. Если критерии окончания поиска удовлетворены, то перейти к этапу 8.

Этап 6. Сформировать новое поколение особей P_{t+1} путем применения квантового вентиля вращения G_{ij} для каждого гена H_{ij} каждой особи H_j текущей популяции P_t: $H_{ij}^{t+1} = G_{ij}^t \cdot H_{ij}^t, i = \overline{1,L}, j = \overline{1,N},$ где H_{ij}^t, H_{ij}^{t+1} – значение i-го гена j-ой особи в текущем и новом поколениях соответственно; G_{ij}^t – матрица преобразования i-го гена j-ой особи. Преобразование G может быть представлено как:

$$G = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix},$$

где θ – угол вращения квантового вентиля G. Квантовый вентиль вращения представляет собой операцию поворота фазы кубита ξ на угол θ против часовой стрелки: $\xi_{ij}^{t+1} = \xi_{ij}^t + \theta_{ij}^t, i = \overline{1,L}, j = \overline{1,N},$ где $\xi_{ij}^t, \xi_{ij}^{t+1}$ – фаза кубита, соответствующего i-му гену j-ой особи в текущем и новом поколениях соответственно; θ_{ij}^t – угол поворота фазы кубита, соответствующего i-му гену j-ой особи на текущей итерации метода. Значение угла вращения фазы кубита определяется как $\theta = k \cdot h(\alpha,\beta),$ где k – коэффициент, определяющий скорость сходимости метода, чем больше k, тем быстрее сходится метод; $h(\alpha, \beta)$ – направление поиска. При задании k учитываются оценки I_j и то, какая часть итераций уже выполнена:

$$k_{ij}^t = I_j \exp(-t/T), \; i = \overline{1,L}, \; j = \overline{1,N}.$$

Направление поиска $h(\alpha, \beta)$ определяется выражением:

$$h(\alpha,\beta) = \begin{cases} 1, d_1 \cdot d_2 \cdot (|\xi_1| - |\xi_2|) > 0, \\ -1, d_1 \cdot d_2 \cdot (|\xi_1| - |\xi_2|) < 0, \end{cases}$$

где $d_1 = \alpha_{ij} \cdot \beta_{ij}$, $d_2 = \alpha_{i,best} \cdot \beta_{i,best}$, $\xi_1 = \operatorname{arctg}(\beta_{ij}/\alpha_{ij})$, $\xi_2 = \operatorname{arctg}(\beta_{i,best}/\alpha_{i,best})$.

Этап 7. Увеличить счетчик итераций: t = t + 1.

Этап 8. В качестве результата работы метода вернуть H_{best} и прекратить поиск.

III. Эксперименты и результаты

Предложенный метод отбора информативных признаков программно реализован и апробирован при решении практических задач технического диагностирования [3]. В результате проведенных экспериментов установлено, что за счет использования априорной информации об обучающей выборке на этапе инициализации начальные решения концентрируются в наиболее перспективных областях пространства поиска, за счет чего обеспечивается существенное ускорение поиска еще на начальном этапе. В процессе поиска решений за счет ускорения операторов стохастического поиска достигается существенное увеличение скорости поиска комбинации признаков, обеспечивающей приемлемый уровень точности диагностической модели. Результаты проведенных экспериментов подтвердили работоспособность программного обеспечения и позволяют рекомендовать предложенный метод для применения на практике при решении задач отбора информативных признаков.

IV. Выводы

В работе решена актуальная задача автоматизации отбора информативных признаков на основе квантовых вычислений.

Научная новизна работы заключается в том, что получил дальнейшее развитие метод отбора информационных признаков, который использует стратегию стохастического поиска для перебора решений и квантовые вычисления для кодирования решений, отличающийся тем, что на этапе генерации начальных решений вероятностно учитываются оценки индивидуальной информативности признаков. Это позволяет повысить скорость отбора признаков.

Практическая ценность работы состоит в том, что создано программное обеспечение, реализующее предложенный метод и позволяющее автоматизировать решение задачи отбора информативных признаков.

Список источников

[1] Субботін С. О. Неітеративні, еволюційні та мультиагентні методи синтезу нечіткологічних і нейромережних моделей: монографія / С. О. Субботін, А. О. Олійник, О. О. Олійник; під заг. ред. С. О. Субботіна. – Запоріжжя: ЗНТУ, 2009. – 375 с.

[2] G. Zhang, L. Hu, and W. Jin, "Quantum Computing Based Machine Learning Method and Its Application in Radar Emitter Signal Recognition" in *Modeling Decisions for Artificial Intelligence*, Berlin: Springer, 2004, pp. 92-103.

[3] Прогрессивные технологии моделирования, оптимизации и интеллектуальной автоматизации этапов жизненного цикла авиационных двигателей : монография / А. В. Богуслаев, Ал. А. Олейник, Ан. А. Олейник, Д. В. Павленко, С. А. Субботин; под ред. Д. В. Павленко, С. А. Субботина. – Запорожье: ОАО "Мотор Сич", 2009. – 468 с.

Lviv Polytechnic National University Institutional Repository http://ena.lp.edu.ua