ПРИМЕНЕНИЕ МАШИНОСТРОЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАССИФИКАЦИИ СИСТЕМ НЕФТЯНЫХ СКВАЖИН

APPLICATION OF MECHANICAL ENGINEERING TO SOLVE THE PROBLEM OF CLASSIFICATION OF OIL WELL SYSTEMS

Zayar Aung
I. Mikhailov
Ye Thu Aung

Summary. The article deals with the application of the data mining method-the support vector machine (SVM) to solve the practical problem of evaluating the efficiency of oil wells. This nonlinear method shows better results than linear regression (LR), which is also a machine learning method. The paper presents and analyzes the principles of solving the classification problem using logistic regression methods and support vector machines. The accuracy of these two algorithms under the same conditions is calculated and compared in experiments.

Keywords: machine learning; data mining; support vector machine; oil wells.

Зеар Аунг
Аспирант, Национальный Исследовательский Университет «Московский Энергетический Институт»
zayaraung53@gmail.com
Михайлов Илья Сергеевич
К.т.н., доцент, Национальный Исследовательский Университет «Московский Энергетический Институт»
fr82@mail.ru
Йе Тху Аунг
Аспирант, Национальный Исследовательский Университет «Московский Энергетический Институт»
yethuaungg55@gmail.com

Аннотация. В статье рассматривается применение метода интеллектуального анализа данных — машины опорных векторов (SVM) для решения практической задачи оценки эффективности нефтяных скважин. Данный нелинейный метод показывает лучшие результаты анализа чем метод линейной регрессии (LR), также являющейся методом машинного обучения. В работе приведены и проанализированы принципы решения задачи классификации с помощью методов логистической регрессии и машины опорных векторов. В экспериментах рассчитаны и сопоставлены точности этих двух алгоритмов при одинаковых условиях.

Ключевые слова: машинное обучение; интеллектуальный анализ данных; машина опорных векторов; нефтяные скважины.

Введение

Развитие цифровизации параметров работы нефтяных скважин, как источников значений параметров для массового производства, так и методов сбора данных в реальном времени, позволяет обеспечивать оптимизацию процесса добычи нефти [1]. Использование машинного обучения для очистки, интеграции, преобразования данных, разработки приложений и оптимизации анализа данных нефтяных скважин является новым научным подходом к решению задачи анализа работы нефтяных скважин. В настоящее время параметры нефтяных скважин, используемые в алгоритме анализа данных, относительно просты, при условии отсутствия параметров, зависящих от других групп параметров, и стандартных способах вычисления оценки данных [2–3]. В статье предлагается нелинейный алгоритм классификации SVM, построение структуры системы разработки данных и модели распознавания полифилетических параметров с использованием SVM через карту пространства признаков высокой размерности и оптимизированную гиперплоскостную классификацию для решения задачи анализа нелинейных параметров нефтяных скважин и распознавания шаблонов совокупностей значений параметров скважин, отражающих их текущее состояние.

1. Полифилетические параметры модели распознавания образов нефтяных скважин

В процессе добычи нефти центр наблюдения собирает, передает, анализирует и выдает в режиме реального времени данные о давлении, температуре, электрическом напряжении, электрическом токе и нагрузке, а также других первичных параметрах, что помогает администратору понимать условия работы нефтяной скважины и обеспечивать её работу в режиме высокой эффективности и низкого потребления [4–5]. Как правило к данным параметрам также относятся пиковые значения электрического тока и напряжения, напорное давление насоса, противодавление, давление масла и давление в затрубном пространстве скважины. Эти данные передаются в автоматизированную систему управления в режиме реального времени. После выполнения линейной аппроксимации и прогнозирования данных, лицо принимающее решение может оценивать состояние скважины в данный момент и прогнозировать её поведение

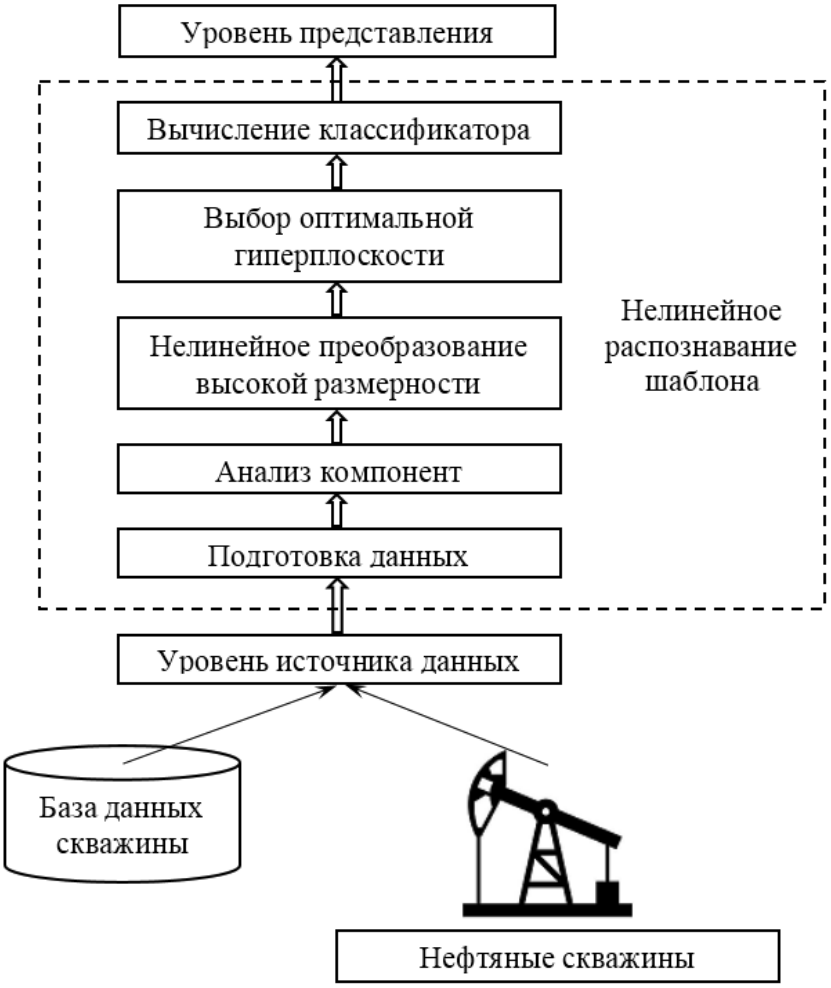


Рис. 1. Модель распознавания шаблона состояния нефтяной скважины

в будущем, для принятия соответствующих компенсирующих управляющих воздействий.

На рисунке 1 представлен процесс распознавания текущей ситуации на скважине.

2. Нелинейный SVM

2.1. Метод ядра

Метод ядра позволяет решить задачу нелинейной классификации с помощью нелинейного преобразования [6]. При условии, что входное пространство является Евклидовым-пространством и признаковое пространство является гильбертовым пространством, метод ядра означает произведение векторов объектов, полученных в процессе преобразования входных данных из входного пространства в признаковое пространство. С помощью метода ядра можно исследовать нелинейные данные с целью получения нелинейного метода SVM. Вся указанная процедура представляет собой работу линейного метода SVM в многомерном пространстве признаков.

Метод ядра показан на рисунке 2.

Общая идея заключается в использовании нелинейного преобразования для изменения входного пространства в пространство признаков, которое может преобразовать модель гиперповерхности в исходном пространстве в гиперплоскость в пространстве признаков. Это означает, что нелинейная задача классификации в исходном пространстве преобразуется в задачу, которая может быть решена линейным SVM в пространстве признаков.

2.2. Метода Опорных Векторов SVM

Общая идея SVM заключается в решении задачи правильной классификации множества данных и максимизации геометрического поля. Может быть несколько

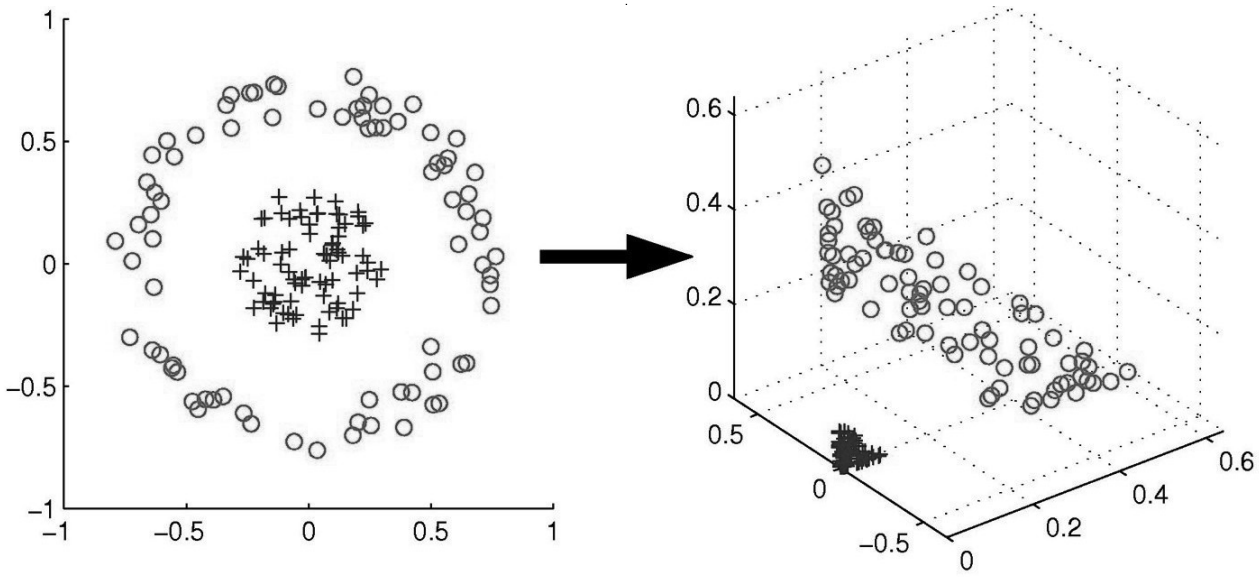
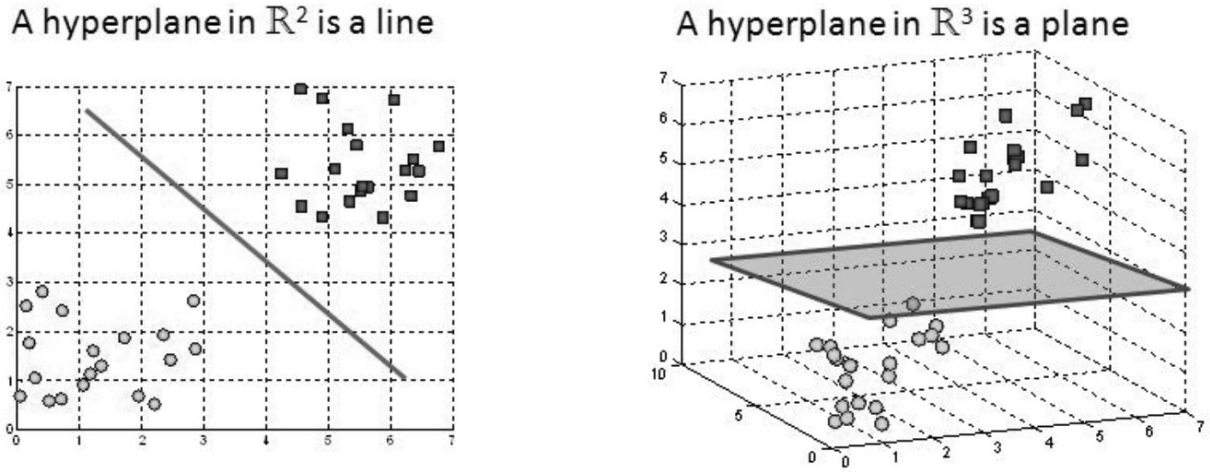


Рис. 2. Использование метода ядра для решения нелинейной задачи



разделяющих гиперплоскостей, но существует только одна разделяющая гиперплоскость с максимальным геометрическим отступом. Прямое объяснение максимизации геометрического поля заключается в том, что гиперплоскость с максимальным геометрическим отступом, полученным из классификации, равна классификации обучающих данных по достаточному фактору определенности. Необходимо не только правильно классифицировать, но и разделять ближайшие точки с достаточным коэффициентом достоверности. Этот процесс может предоставить определенные данные с хорошей прогностической способностью, которая называется способностью обобщения.

При решении нелинейной задачи после преобразования в многомерное пространство, как правило, трудно найти гиперплоскость, которая может полностью разделить точки данных, а это значит, что есть некоторые особые точки. Но после удаления этих особых точек большая часть точек становится линейно разделима. Чтобы решить эту проблему, мы импортируем скользящую переменную в обучающую выборку. В ситуации мягких краёв задача обучения SVM будет иметь вид:

$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + C\sigma_{i\ \ 1}\xi_i. \quad (1)$$

$$\text{s.t. } y_i(wx_i+b) \geq 1-\xi_i. \quad (2)$$

Где C-параметр штрафа. При увеличении C также увеличивается штраф за ошибки классификации. Необходимо отрегулировать целевую функцию, чтобы мини-

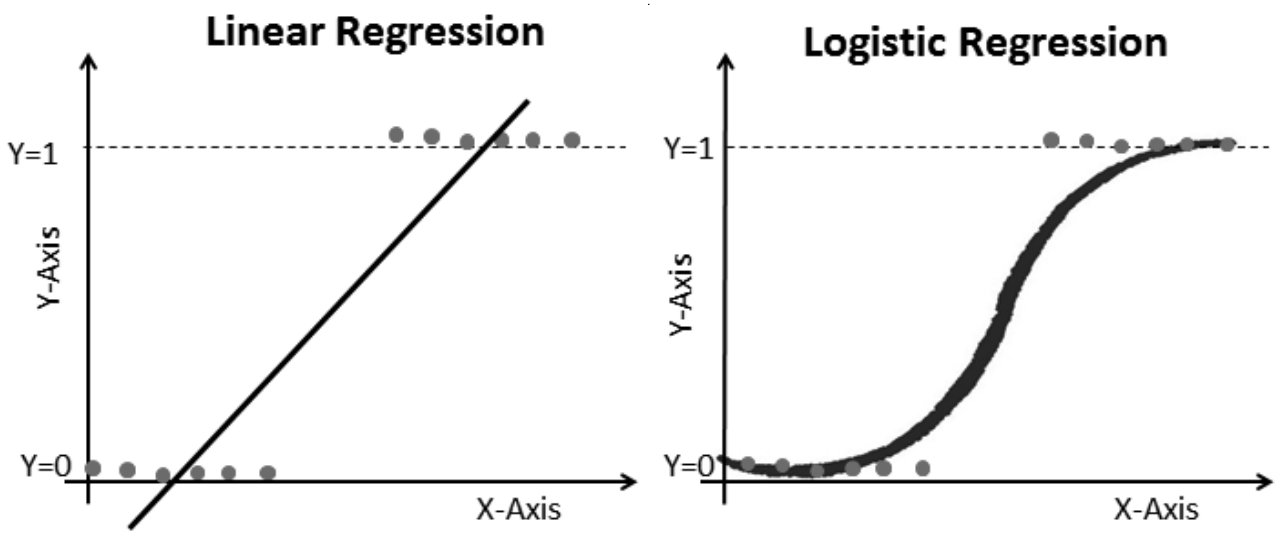


Рис. 3. Функция распределения логистической регрессии и функция плотности

мизировать количество особых точек, одновременно максимизируя отступ от гиперплоскости.

3. Алгоритм линейной логистической регрессии

Алгоритм линейной логистической регрессии — это классический метод классификации в исследовании статистики, относящийся к линейной логарифмической модели. Эта модель классификации представляет собой условное распределение вероятностей P (Y / X), которое является моделью суждения. Она может быть получена из модели линейной регрессии hw (x)= $_{W}{}^{T}{}_{X}$ и сигмоидной кривой:

$$P(Y=1|X) = \frac{1}{1+e^{w}}. \quad (3)$$

Где x-вход, y-выход, w-взвешенный коэффициент и wx — внутреннее произведение. Функция распределения логистической регрессии и функция плотности показаны на рисунке 3.

Логистическая регрессия сравнивает разницу между двумя условными вероятностями и относит обучающий пример x в большую вероятностную группу. Для обучающего набора данных можно использовать функцию максимального правдоподобия для оценки параметров модели для получения логистической модели. Вводятся следующие предположения.

$$P(Y=1x) = f(x),\ P(Y0=x)\ 1 = f(x). \quad (4)$$

Функция правдоподобия имеет вид:

$$\prod_{i=1}^{N}[f(x)]\ [1-f(x)]1-\text{yi}. \quad (5)$$

Логарифмическая функция правдоподобия имеет вид:

$$L(w)=\sum_{i=1}^{N}[y\ log\ f(x)+(1-y)log(1-\ f(x))]. \quad (6)$$

4. Реализация и результаты эксперимента

4.1. Эксперимент по оценке эффективности планирования работы нефтяной скважины

Эффективность системы — это наиболее важный фактор качества работы системы добычи. Эффективность системы добычи — это отношение полезного количества добытой жидкости к потребляемой мощности в единицу времени, что является существенным фактором производства. В результате эксперимента в качестве целевого фактора была выбрана эффективность системы. Предполагается, что значение эффективности системы выше 45% является положительным, меньше 45% — отрицательным.

В интеллектуальном анализе данных такие параметры, как нагрузка, температура и электрическое напряжение насоса, подходят для решения задачи классификации в модели оценки. При анализе эффективности насосной системы рассматриваются влияющие на неё факторы, перечисленные в таблице 1. Данные, приведенные в таблице 1, были получены для каждой нефтяной скважины в одно время.

Для улучшения результатов выполненной работы в соответствии с полученными данными были выполнены следующие действия.

Таблица 1. Параметры нефтяной скважины

Параметры	Единица Измерения	Параметры	Единица Измерения
Глубина	[м]	Реактивная мощность	[КВ]
Период работы	[ч]	Давление масла	[МПа]
Максимальная нагрузка	[КН]	Максимальное давление	[МПа]
Минимальная нагрузка	[КН]	Минимальное давление	[МПа]
Коэффициент мощности	[1]	Давление продукции	[МПа]
Активная мощность	[КВ]	Напряжение	[В]
Максимальная активная мощность	[КВ]	Ток	[А]

Таблица 2. Параметры добычи нефти в скважине

Параметры	Единица Измерения	Параметры	Единица Измерения
Потребление Жидкости	[м3 / день]	Доплеровская скорость (массив)	[Герц]
Потребление Газа	[м3 / день]	Газовая пустотная фракция (массив)	[%]
Обводненность	[%]	Скорость звука	[М/С]
Температура	[˚C]	давление жидкости	[МПа]

Таблица 3. Результаты классификации

№	Реальное значение	Прогноз LR	Прогноз SVM	№	Реальное значение	Прогноз LR	Прогноз SVM
1	0	1	0	16	0	0	1
2	0	0	0	17	0	0	0
3	0	0	0	18	0	0	0
4	0	0	0	19	0	0	0
5	0	0	0	20	0	0	0
6	0	1	1	21	1	1	1
7	1	0	1	22	0	1	0
8	1	1	1	23	0	0	0
9	0	1	1	24	0	0	0
10	1	1	1	25	1	1	1
11	1	0	0	26	0	1	1
12	1	0	0	27	0	1	1
13	0	0	1	28	0	1	1
14	0	1	1	29	1	0	0
15	0	1	1	30	1	0	0

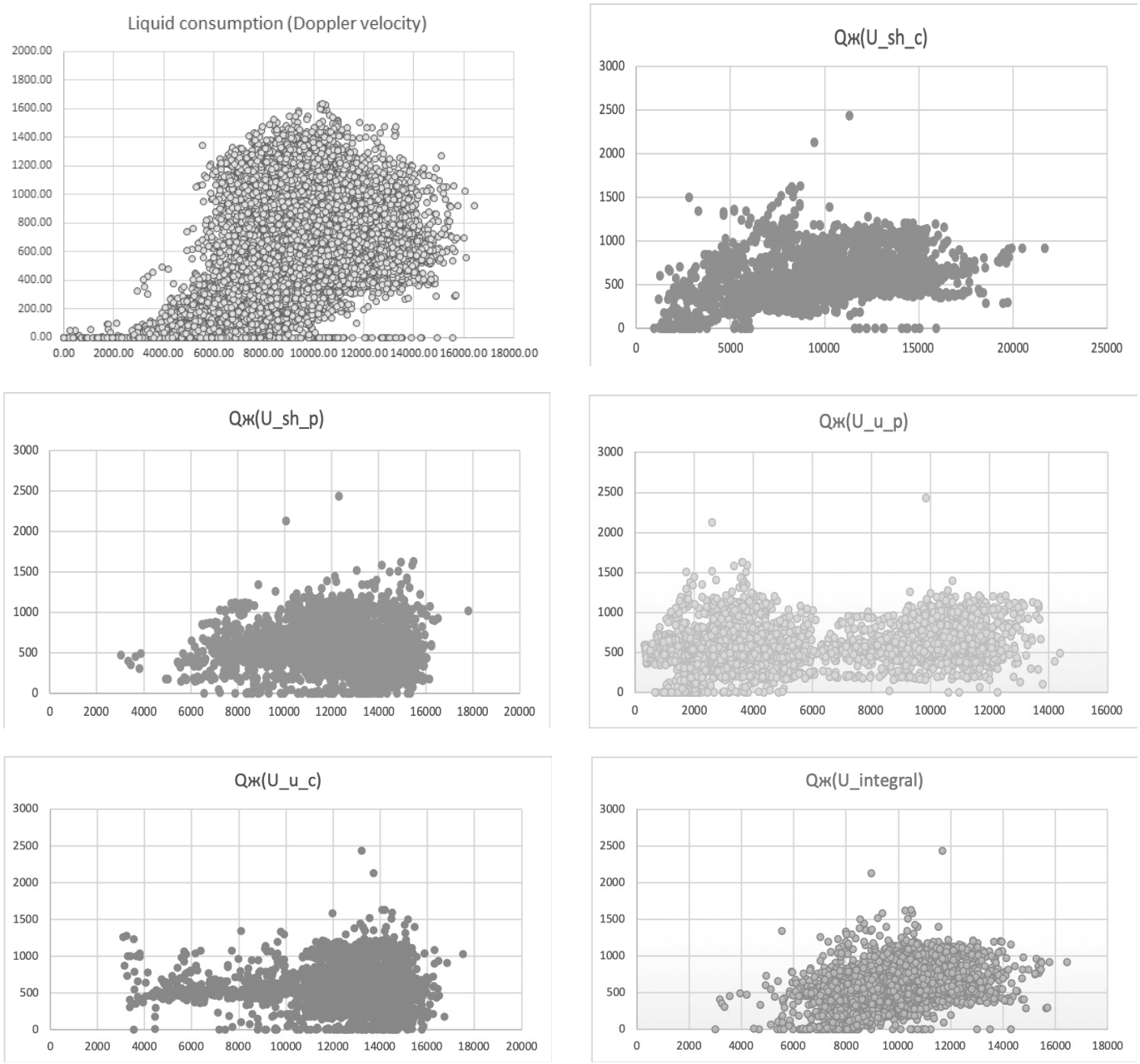


Рис. 4. Зависимость расхода жидкости от интегральной доплеровской скорости

1. С целью повышения эффективности данных была собрана вся возможная относящаяся к ним информация.
2. Была выполнена предварительная обработка данных методами сглаживания, нормализации и шумоподавления.
3. Создана модель оценки для решения реальных задач.
4. Выполнена оценка полученных моделей.
5. Выработана оптимальная модельная схема.
6. Выполнено сравнение результатов с реальными данными, после чего выполнено обновление модели.

4.2. Результаты классификации

Эксперимент проводился на языке python на примере нефтяных скважин месторождения с использованием алгоритмов SVM и LR. 1980 нефтяных скважин были выбраны в качестве обучающего множества, оставшиеся 30 нефтяных скважин в качестве тестовой выборки. Согласно опыту, параметр штрафа C был установлен 0.8, функция оценки RBF и стандартное отклонение 0.5 для модели SVM; параметр штрафа C=1 для LR. Сравнение прогнозируемой и реальной эффективности приведено в таблице 3.

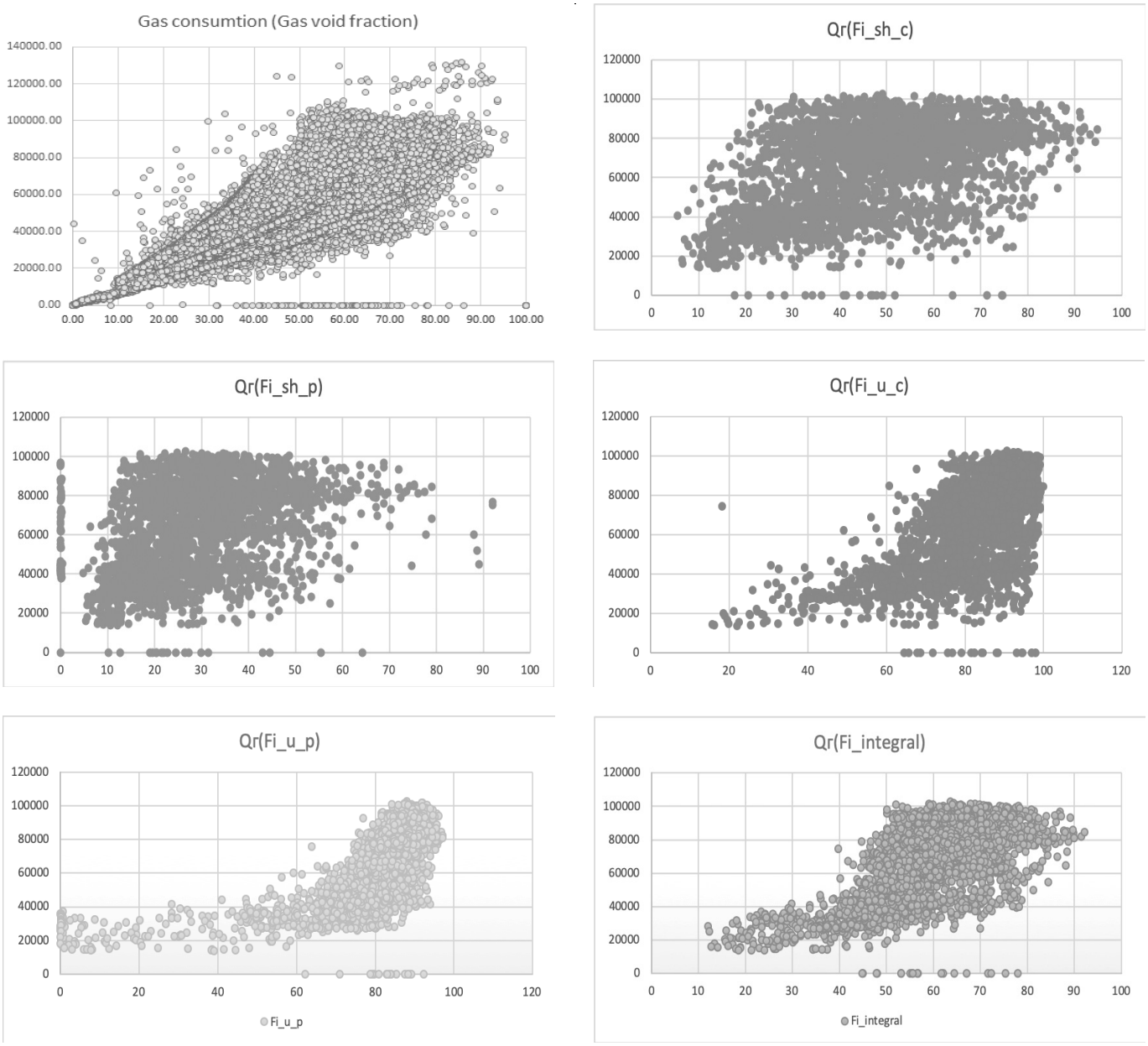


Рис. 5. Зависимость между расходом газа и интегральной долей газовых пустот

При использовании логистической модели было обнаружено 21 правильных результатов классификации, что означает, что точность достигает 75%. В рамках модели SVM найдено 18 правильных классификаций с точностью 90%, которые удовлетворяют условиям прогнозирования. С помощью метода уменьшения размерности PCA можно уменьшить размерность данных 17 до 2 с учетом визуализации, результат которой показан на рисунке 4. Множество точек на рис. 4 означает определенный набор данных. Квадраты означают правильную классификацию SVM, а звезды — LR. Перекрывающиеся части корректны в обоих алгоритмах, а красные крестики являются ошибками классификации.

5. Результаты эксперимента SVM и LR

В предметной области нефтяных скважин распределение данных осложнено высокой размерностью пространства данных, что может оказать большое влияние на сбор первичных данных. В этой ситуации возможна ошибка сбора одного или нескольких видов данных, а также неравномерное распределение данных. Классический ручной анализ, такой как применение диаграмм, линейный анализ или логистическая регрессия, не позволяет достигнуть высокого качества классификации. В этом случае машина опорных векторов с использо-

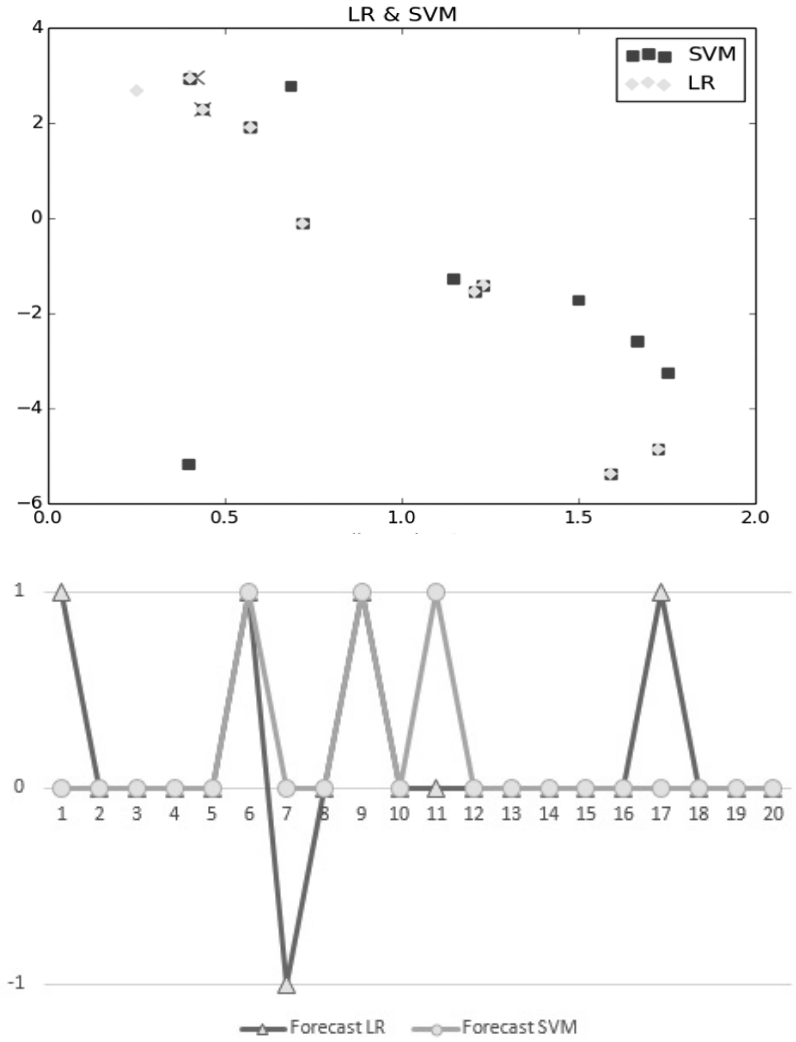


Рис. 6 Результаты эксперимента SVM и LR

ванием метода ядра лучше подходит для нелинейного сложного процесса обработки данных.

6. Заключение

В работе выполнен теоретический анализ метода опорных векторов и метода логистической регрессии. Показано, что нелинейный алгоритм SVM работает лучше, чем линейный алгоритм LR при анализе системы нефтяных скважин и прогнозировании их эффективности. В дальнейшем необходимо разработать метод классификации на базе машины опорных векторов, позволяющий осуществлять классификацию исходного множества данных на несколько классов с возможностью оценки степени близости к каждому из этих классов.

ЛИТЕРАТУРА

1. Yong Soo Kim. Performance evaluation for classification methods: A comparative simulation study[J]. Expert Systems With Applications, 2009,373.
2. Hanuman Thota, Raghava Naidu Miriyala, Siva Prasad Akula, K. Mrithyunjaya Rao, Chandra Sekhar Vellanki, et al. Performance Comparative in Classification Algorithms Using Real Datasets[J]. Journal of Computer Science & Systems Biology, 2009, 0201.
3. HungLinh Ao, Junsheng Cheng, Yu Yang, Tung Khac Truong. The support vector machine parameter optimization method based on artificial chemical reaction optimization algorithm and its application to roller bearing fault diagnosis. Journal of Vibration and Control.2015(12).
4. Rimjhim Agrawal, Thukaram Dhadbanjan. Identification of Fault Location in Distribution Networks Using Multi Class Support Vector Machines. International Journal of Emerging Electric Power Systems.2012(3).

5. Snehal A. Mulay, P. R. Devale, G. V. Garje. Intrusion Detection System Using Support Vector Machine and Decision Tree. International Journal of Computer Applications.2010(3).
6. Wang Liejun, Lai Huicheng, Zhang Taiyi. An Improved Algorithm on Least Squares Support Vector Machines. Information Technology Journal.2008(2).
7. R. Cogdill, P. Dardenne. Least-squares support vector machines for chemometrics: an introduction andevaluation. Journal of Near Infrared Spectroscopy.2004(2).
8. Ke Lin, Anirban Basudhar, Samy Missoum. Parallel construction of explicit boundaries using support vector machines. Engineering Computations.2013(1).
9. Ashkan Moosavian, Hojat Ahmadi, Babak Sakhaei, Reza Labbafi. Support vector machine and K-nearest neighbour for unbalanced fault detection. Journal of Quality in Maintenance Engineering.2014(1).
10. Long Zhang, Jianhua Wang. Optimizing parameters of support vector machines using team-search-based particle swarm optimization. Engineering Computations. 2015(5).
11. Bashmakov, A. I., Bashmakov, I. A. Intellectual information technologies. Benefit // –M.: Publishing MGTU im. N. Uh. Bauman, 2005.
12. Вапник В. Н., Статистическая теория обучения: Нью-Йорк: John Wiley \ & Sons, 1998, 740 С.
13. RRDtool. URL: http://oss.oetiker.ch/rrdtool/ (дата обращения: 25.04.2013).
14. Kaufman L., Rousseeuw P. J., Нахождение групп в данных введение в кластерный анализ: NJ, Hoboken, USA: John Wiley \ & Sons, 2005, 355 p.
15. Scholkopf Б., С. Платт Дж., Shawe-Тейлор Дж., Смола А. Дж., Уильямсон К. К., Нейронные вычисления, 2001, том. 13, С. 1443–1471.
16. Лин ХС.— Ти, Линч.— J., Weng R. C., Машинное обучение, 2007, Vol. 68, PP.
17. Санчес-Фернандес М., Арен-Гарсия Дж., Перес-Крус Ф., Сделки IEEE по обработке сигналов, 2004, вып. XX, нет. V, PP.
18. Черкасских В., M. GPE фирмы Intel 2415, Спрингер-Верлаг, Берлин-Хайдельберг, 2002, с. 687–693.
19. Ма Я., С. Перкинс, Тез. Доклад на 9-м симпозиуме ACM'03, Вашингтон, округ Колумбия, США, 2003, с. 613–618.
20. Yeon Su Kim. Evaluation of the effectiveness of classification methods: comparative modeling. Expert systems with applications, 2009, 373 p.
21. Hanuman Thota, Raghava Miriyala, Siva Prasad Akula, K. Mrithyunjaya RAO, Chandra Sekhar Vellanki, et al. Performance comparison in classification algorithms using real data sets. Journal of computer science and systems biology, 2009, 02–01.

© Зеар Аунг (zayaraung53@gmail.com),
Михайлов Илья Сергеевич (fr82@mail.ru), Йе Тху Аунг (yethuaungg55@gmail.com).
Журнал «Современная наука: актуальные проблемы теории и практики»

Национальный Исследовательский Университет «Московский Энергетический Институт»