УДК 004.891.3

© 2016 г. **А.П. Саенко**
(Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики)

ПРОЕКТИРОВАНИЕ НЕЧЕТКОГО КЛАССИФИКАТОРА ДЛЯ ИДЕНТИФИКАЦИИ МИКРООРГАНИЗМОВ НА ПОЛУТОНОВЫХ ЦИФРОВЫХ ИЗОБРАЖЕНИЯХ

В статье рассмотрено проектирование нечеткого классификатора для идентификации и обнаружения вредоносных бактерий в биологическом материале. Предлагаемый алгоритм показывает достаточно высокую эффективность по сравнению с другими классификаторами.

Ключевые слова: распознавание образов, нечеткая логика.

DOI: 10.22250/isu.2016.49.97-104

Введение

В последние годы методы обработки цифровых изображений и машинного обучения все чаше применяются для решения большого количества задач в различных отраслях науки и техники, позволяя решать широкий спектр прикладных задач обнаружения различных объектов на изображениях.

С другой стороны, одной из актуальных проблем в области биомедицины является своевременное выявление вредоносных микроорганизмов в биологическом материале. Несмотря на то, что существуют отечественные (ГОСТ 26668-85, ГОСТ 26669-85, ГОСТ Р 51448-99, ГОСТ 10444.15-94 и др.) и международные (ISO 4831, ISO 4832, ISO 4833, ISO 6579 и др.) стандарты, определяющие методы выявления микроорганизмов, они не лишены недостатков и в общем случае состоят из нескольких этапов (рис. 1). В начале осуществляются отбор проб и их подготовка к анализу, а также приготовление реагентов и питательных сред. Затем пробы помещаются в селективную питательную среду и инкубируются в определенных условиях в течение некоторого времени. После этого проводятся серологический или биохимический анализы.

В целом, оценка риска заражения биологического материала путем определения вида и количества микроорганизмов в пробах занимает до 5 дней, что может быть неприемлемо, например, при исследовании скоропортящихся продуктов питания, а также предъявляет значительные требования к оснащению лаборатории и квалификации персонала.

Частично указанные недостатки решает недавно предложенный метод анализа, основанный на флуоресцентной микроскопии [1].

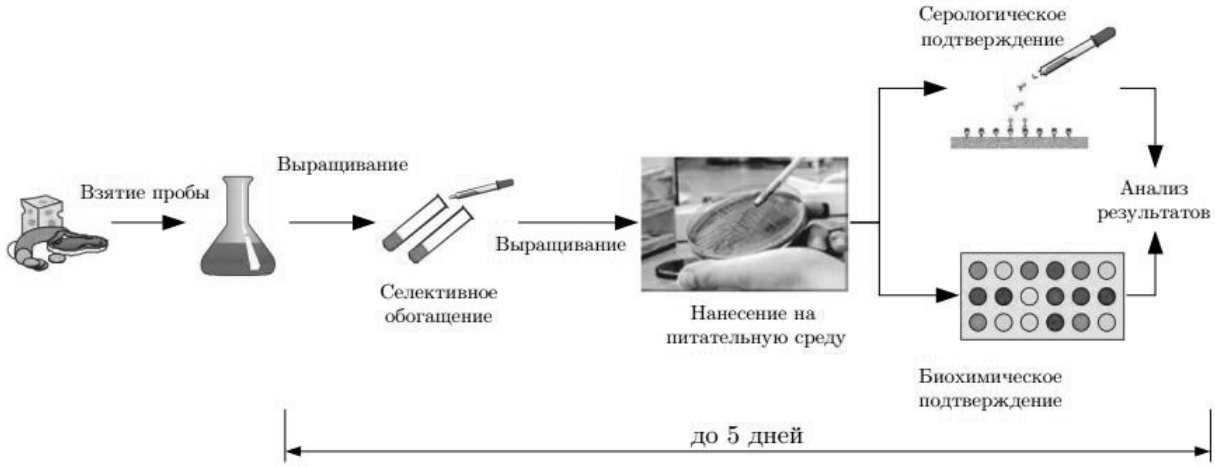

Рис. 1. Стандартная процедура выявления микроорганизмов.

Его реализация на основе мобильной аналитической платформы (рис. 2) работает следующим образом. Проба помещается в специальную мембрану, представляющую собой дискообразную полость диаметром 10 мм и толщиной 2 мм.

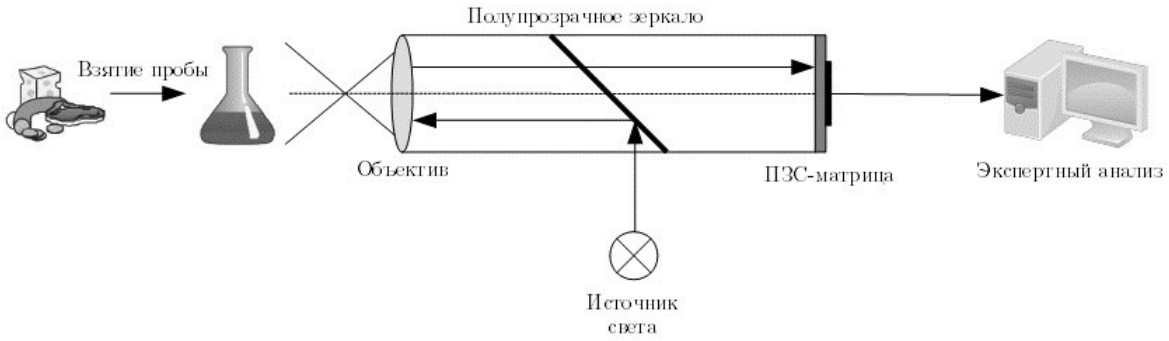

Рис. 2. Функциональная схема мобильной аналитической платформы.

Сама мембрана при этом располагается на подвижной платформе, которая осуществляет прецизионные перемещения в горизонтальной и вертикальной плоскостях. Источник света, подвергая мембрану излучению определенной длины волны, возбуждает флуоресценцию, фотографические монохроматические изображения которой записываются камерой. В итоге для каждой пробы образуется набор цифровых полутоновых изображений, которые затем анализируются экспертом с целью обнаружения микроорганизмов. Таким образом, данная платформа значительно сокращает время анализа (до нескольких часов), однако по-прежнему требует наличия высококвалифицированного персонала.

Дальнейшее совершенствование методов выявления микроорганизмов в биологическом материале остается актуальной задачей, решение которой может значительно ускорить и удешевить процесс анализа. Ранее на примере обнаружения бактерий вида *Legionella pneumophila* (палочковидный возбудитель легионеллеза длиной около 2 мкм и шириной 0,3–0,9 мкм) показано, что внедрение классических методов обработки цифровых изображений и распознавания образов не всегда достаточно эффективно из-за изменчивости наблюдаемых объектов и низкого контраста исходных изображений, а также характеризуется низкой степенью интерпретируемости экспертами в предметной области [2, 3]. Это делает целесообразным применение теории нечеткой логики для решения поставленной задачи.

Задача идентификации объектов

В соответствии с классическим алгоритмом обработки и анализа изображений [4, 5] полученные изображения проходят этапы предварительной обработки (улучшения) и сегментации. В результате образуется набор фрагментов изображений объектов, потенциально представляющих интерес (рис. 3).

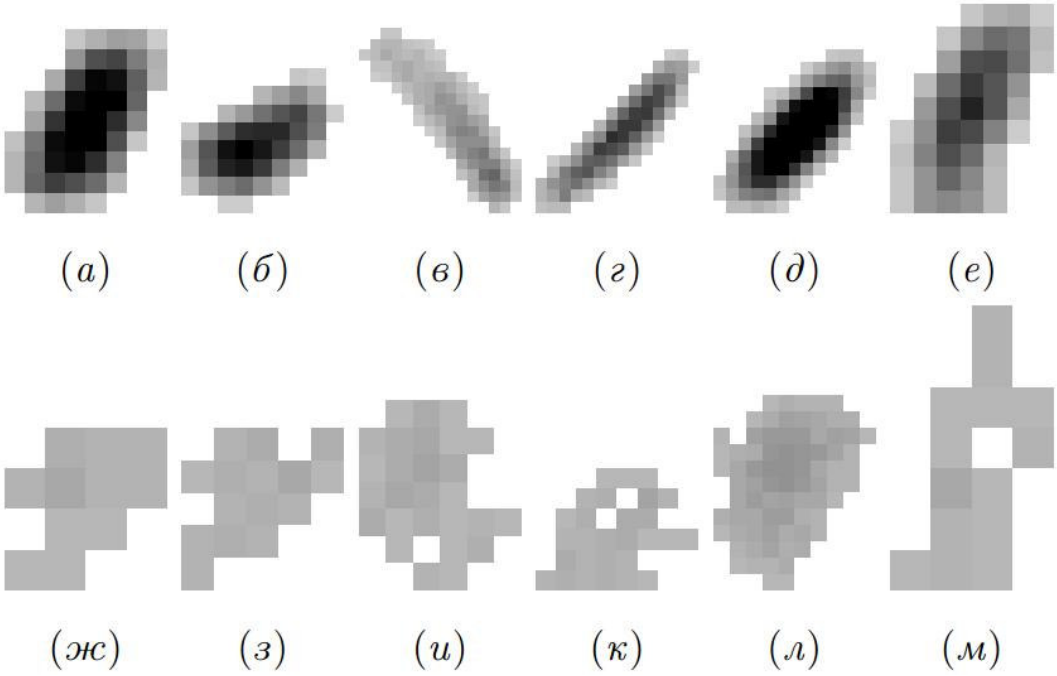

Рис. 3. Примеры сегментированных объектов: бактерий (а-е) и сторонних артефактов (ж-м).

Таким образом, задача обнаружения бактерий сводится к задаче классификации на два непересекающихся класса («бактерия» и «сторонний артефакт») и по сути является частным случаем задачи машинного обучения, которая в общем виде заключается в необходимости при конечном множестве классов $Y = 1,2,...,l$ построить алгоритм, который по объекту x определяет точное или достаточно точное значение $y(x)$ [5, 6]. В качестве исходных данных принимается пространство допустимых объектов X, пространство меток Y, а также целевая функция $y(x)$, заданная в конечном множестве точек обучающей выборки $y(x_1)$, $y(x_2)$, …, $y(x_m)$.

Обучающая выборка представляет собой матрицу с описанием объектов $\mathbf{X}$ и вектор меток $\mathbf{Y}$:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}, \boldsymbol{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix},$$

где m – количество объектов; n – количество признаков этих объектов. Таким образом, каждый ряд матрицы $\boldsymbol{X}$ соответствует одному объекту x_i, представленному в виде вектора признаков, а каждый элемент $y_i \in \{0,1\}$ определяет класс i-го объекта.

Функция потерь $L(A(x), y(x))$ показывает, насколько ответ $A(x)$ соответствует верному ответу $y(x)$, и определяется как:

$$L(A(x), y(x)) = \begin{cases} 1, & A(x) \neq y(x), \\ 0, & A(x) = y(x). \end{cases}$$

Соответственно алгоритмы машинного обучения должны решать задачу оптимизации в виде:

$$\frac{1}{m}\sum_{i=1}^{m} L(A(x), y(x)) \to \min.$$

Оценить сложность задачи классификации можно по рис. 4., на котором представлена визуализация выборки методом главных компонент [7].

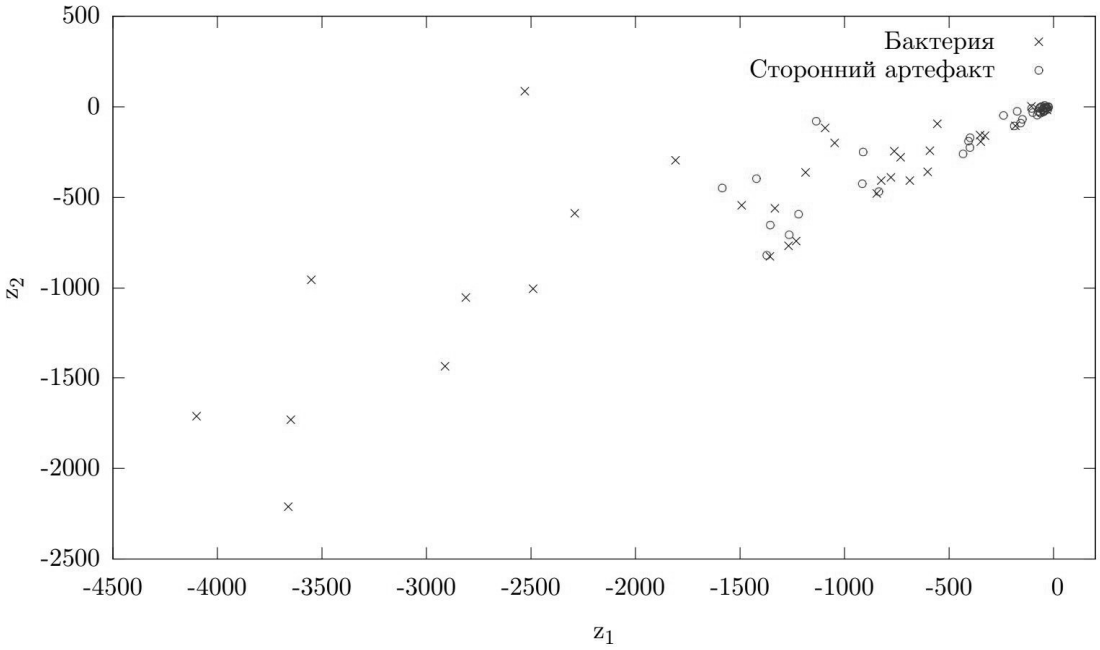

Рис. 4. Визуализация выборки.

В ходе численного эксперимента анализировались 30 различных признаков, рассчитываемых с помощью программного пакета Halcon. В результате установлено, что релевантными являются нормированные масштабные моменты 2-го порядка по строке и столбцу, нормированный масштабный относительный момент, выпуклость, неделимость, радиус внутренней окружности, прямоугольность и округленность [11].

Проектирование и оценка эффективности нечеткого классификатора

Теория нечеткой логики расширяет классическое понятие множества, допуская, что функция принадлежности элемента множеству может принимать любые значения из интервала $[0;1]$ [9]. Принято считать, что нечетким классификатором является любой классификатор, использующий теорию нечетких множеств [10], для которого справедливо

$$\sum_{i=1}^{l} \mu_i(s) = 1,$$

где $\mu_i(s)$ – степень принадлежности объекта s к i-му классу.

Эмпирические функции принадлежности для каждого признака объекта строятся исходя из его нормализованной гистограммы, а затем экстраполируются в одну из стандартных функций (например, в сигмоидные – рис. 5). База нечетких правил содержит правила из табл. 1.

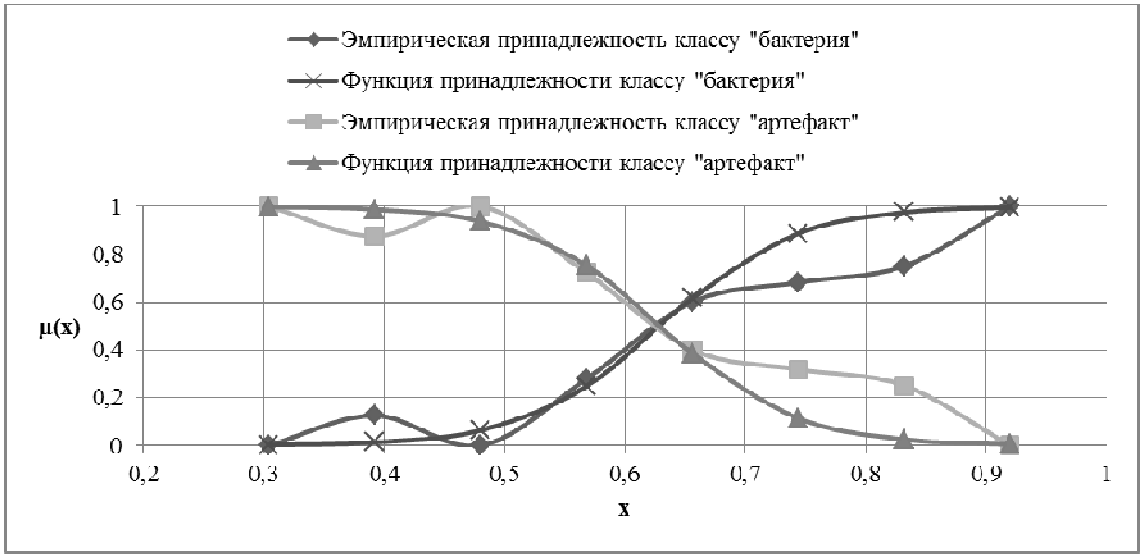

Рис. 5. Эмпирические и экстраполированные функции принадлежности для признака «прямоугольность».

Таблица 1

Признаки								Класс
Момент по строке	Момент по столбцу	Относительный момент	Выпуклость	Неделимость	Радиус	Прямоугольность	Округленность	
0 – признак не участвует в правиле, 1 – бактерия, 2 – артефакт								
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2
0	0	1	1	1	1	1	1	1
0	0	0	1	1	1	1	1	1
1	1	1	0	1	0	1	1	1
1	1	1	1	1	1	0	1	1
0	0	2	2	2	2	2	2	2
0	0	0	2	2	2	2	2	2
2	2	2	0	2	0	2	2	2
2	2	2	2	2	2	0	2	2

Обычно оценка эффективности алгоритмов классификации из-за существенной неформальности большинства задач распознавания производится экспериментально и выражается в способности выбранных алгоритмов принимать верные решения, что характеризуется уровнем ошибок первого («ложный пропуск» – событие ложно не обнаруживается) и второго рода («ложное обнаружение» – событие ошибочно считается произошедшим).

Если количество объектов для каждого класса в тестовом наборе

$$N = Np + Nn,$$

где N – общее количество объектов; Np – количество бактерий; Nn – количество сторонних артефактов, а количество ложных пропусков FN и ложных обнаружений FP, то количество верных пропусков и верных обнаружений определяется как

$$TP = Np - FN,$$

$$TN = Nn - FP.$$

При этом уровни ошибок выражаются следующим образом:

$$nFN = \frac{FN}{Np} \cdot 100\%,$$

$$nFP = \frac{FP}{Nn} \cdot 100\%,$$

$$nTN = \frac{TN}{Nn} \cdot 100\%,$$

$$nTP = \frac{TP}{Np} \cdot 100\%.$$

Одним из способов оценки эффективности алгоритмов классификации в заданных условиях является мера расстояния до точки (0, 1) на ROC-диаграмме (от англ. Receiver Operating Characteristic – операционная характеристика приемника), которая вычисляется в виде [8]:

$$E = \sqrt{FP_{rate}^2 + (1 - TP_{rate})^2}.$$

При этом минимальное возможное значение 0 соответствует наилучшей эффективности ($FP_{rate} = 0$, а $TP_{rate} = 1$, т.е. все бактерии верно определены как экземпляры класса «бактерия» и ни один сторонний артефакт не определен как экземпляр класса «бактерия»). Максимальное значение $\sqrt{2}$ отображает наихудшую эффективность при $FP_{rate} = 1$ и $TP_{rate} = 0$.

Таким образом, для процентного выражения эффективности классификатора, принимая за 100% максимальную эффективность [2], получаем:

$$E_{\%} = \left(1 - \frac{E}{\sqrt{2}}\right) \cdot 100\%.$$

Сравнение предлагаемого нечеткого классификатора с классическими алгоритмами приведено в табл. 2.

Таблица 2

Классификатор	Реальный класс	Распознаны как		Эффективность
		бактерии	артефакты	
Нечеткий классификатор	бактерии	57	3	92,09%
	артефакты	6	54	
Метод опорных векторов	бактерии	50	10	82,48%
	артефакты	11	49	
Случайный лес	бактерии	53	7	82,60%
	артефакты	13	47	
Дерево решений	бактерии	46	14	72,19%
	артефакты	19	41	
k-ближайших соседей	бактерии	45	15	72,39%
	артефакты	18	42	
Метод Байеса	бактерии	55	5	38,43%
	артефакты	52	8	

Заключение

В работе показаны недостатки существующих методов выявления вредоносных микроорганизмов в биологическом материале, показаны способы их совершенствования, заключающиеся в дальнейшей автоматизации процесса. В частности, описан принцип действия мобильной аналитической платформы и предложен нечеткий классификатор для мгновенного обнаружения бактерий в продуктах питания. Приведен сравнительный анализ эффективности различных алгоритмов классификации, показавший высокую эффективность и перспективность дальнейшего применения теории нечеткой логики для решения подобного рода задач.

ЛИТЕРАТУРА

1. *Lerm S., Holder S., Gopfert A., Futterer R., Linss G.* Concepts of a scanning hardware platform for high-resolution image processing with Lab-on-a-chip analysis // The Proceedings of the 15th International Symposium MECHATRONIKA. – Prague, 2012. – P.1-4.
2. *Саенко А.П., Мусалимов В.М., Лерм Ш., Линц Г.* Применение методов машинного обучения для обнаружения бактерий в продуктах питания // Научно-технический вестник информационных технологий, механики и оптики. – 2014. – №1 (89) – С.93-98.
3. *Саенко А.П.* Оценка эффективности обнаружения бактерий методами обработки цифровых изображений и интеллектуального анализа данных // Сборник «Фундаментальные и прикладные проблемы надежности и диагностики машин и механизмов: одиннадцатая сессия международной научной школы». – СПб.: ИПМаш РАН, 2013. – С.318-321.
4. *Латыев С.М., Воронин А.А., Андинг К., Линц Г., Курицын П.А.* Оптико-электронные методы и средства идентификации веществ и материалов // Известия вузов. Приборостроение. – 2013. – Т. 56, № 10. – С.81-87.
5. *Saenko A., Musalimov V., Lerm S.* Analytical imaging of bacteria in foodstuff // Proceedings of the 10th IEEE International Conference on Communications. Bucharest, 2014. – P.135-138.
6. *Дьяконов А.Г.* Анализ данных, обучение по прецедентам, логические игры, системы WEKA, RapidMiner и MatLab (Практикум на ЭВМ кафедры математических методов прогнозирования): Учебное пособие. – М.: МГУ, 2010.
7. *Pearson K.* On lines and planes of closest fit to systems of points in space // Philosophical Maga-

zine. – 1901. – № 2 (11). – P.559-572.
8. *Bramer M.* Principles of data mining. – 2nd ed. – Springer, 2013.
9. *Zadeh L.A.* Fuzzy Sets // Information and Control. – 1965. – Vol. 8. – P.338-353.
10. *Kuncheva L.I.* Fuzzy Classifier Design // Of Studies in Fuzziness and Soft Computing. – Springer, 2000. – Vol. 49.
11. HALCON Version 11.0.1 – HALCON / HDevelop Reference Manual. MVTec Software GmbH, 2012.

Статья представлена к публикации членом редколлегии Е.Л. Ереминым.

E-mail:
Саенко Алексей Петрович – alexey.saenko@gmail.com.

МЕЖДУНАРОДНАЯ IEEE-ЕВРАЗИЙСКАЯ КОНФЕРЕНЦИЯ ПО ЭНЕРГЕТИКЕ, приуроченная к международной выставке ASTANA EXPO–2017, МЕЖДУНАРОДНАЯ IEEE-СИБИРСКАЯ КОНФЕРЕНЦИЯ ПО УПРАВЛЕНИЮ И СВЯЗИ,

КАЗАХСТАН, г. АСТАНА, 29–30 ИЮНЯ 2017 г.

http://sibcon.sfu-kras.ru

Международная IEEE-Евразийская конференция по энергетике, приуроченная к международной выставке ASTANA EXPO-2017 и тринадцатая IEEE-Сибирская конференция, посвященная достижениям в области разработки и создания систем управления и связи, проводится с 29 по 30 июня 2017 года в г. Астане, Казахстан, на базе Казахского агротехнического университета им. С. Сейфуллина. Конференция регулярно организуется красноярской и томской группами и студенческим отделением IEEE, компанией National Instruments для того, чтобы поддерживать междисциплинарные дискуссии и взаимодействие среди ученых и инженеров, развивать международное сотрудничество через участие в деятельности профессиональных сообществ Института IEEE.

ТЕМЫ

1. Фундаментальные проблемы теории управления и связи.
2. Энергосбережение и энергетика будущего.
3. Компьютерные измерительные технологии, датчики и системы.

Программа конференции предусматривает заседания секций с устными докладами, специальные заседания, краткие курсы и культурную программу.