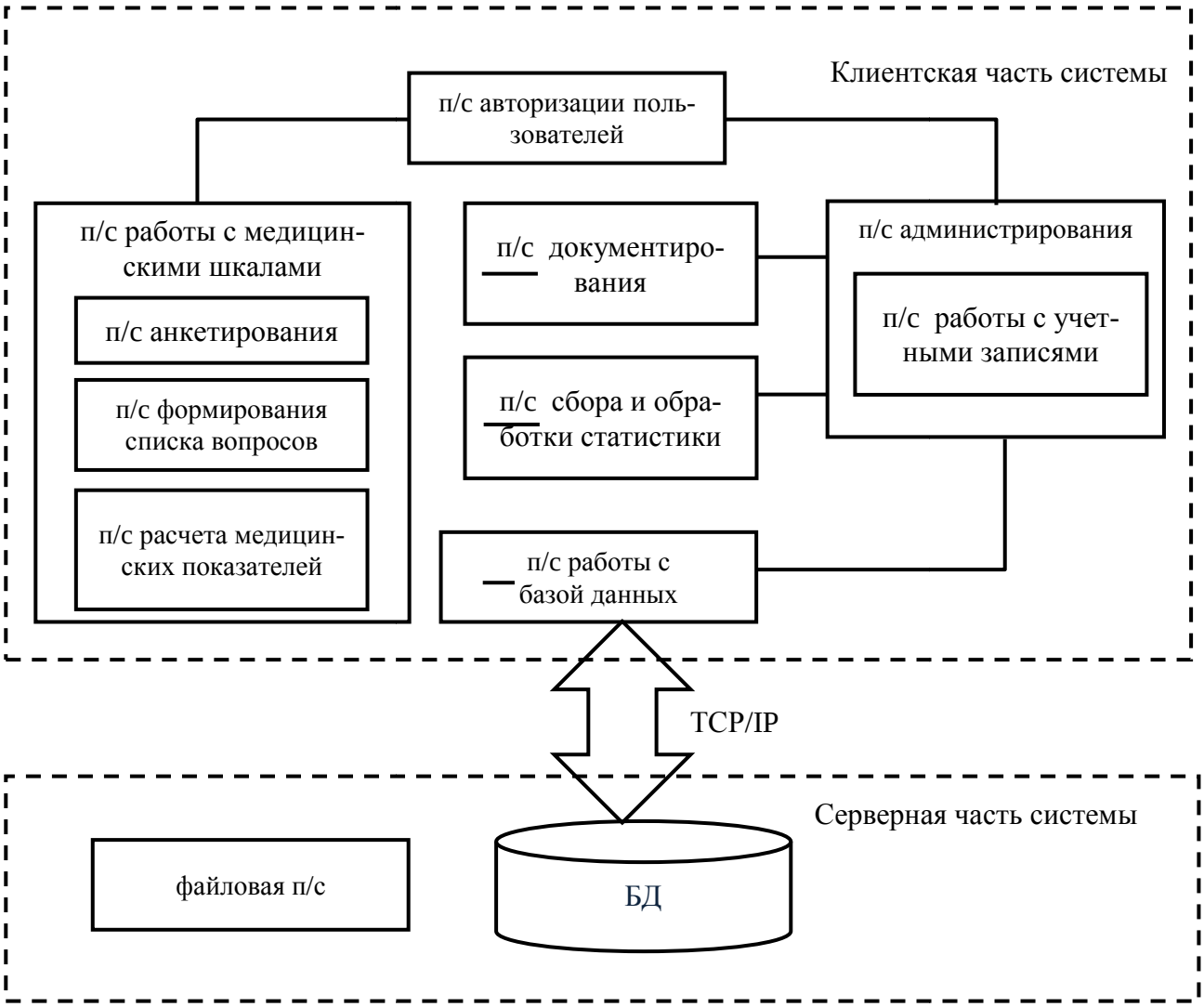

Рис. 2. Структурная схема системы

О.Б. Рузибаев, Ш.Б. Сайфуллаев, Д.А. Алиева

НЕКОТОРЫЕ МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ ДЛЯ ДИАГНОСТИКИ ЗАБОЛЕВАНИЙ РАКА МОЛОЧНОЙ ЖЕЛЕЗЫ

(Ташкентский университет информационных технологий)
(Республиканский онкологический научный центр (РОНЦ МЗ РУз))

В последнее время наблюдается тенденция роста и широкого распространения Рака молочной железы (РМЖ) среди женщин в возрастной группе 35-55 лет. Диагноз РМЖ является широко обсуждаемой и глобальной проблемой, в связи с чем, ранняя диагностика становится актуальной проблемой здравоохранения Республики Узбекистан. Раннее выявление РМЖ имеет большое значение для спасения жизней, позволяет свести к минимуму риск распространения заболевания ткани в другие органы. Точные и надежные методы, необходимые

для раннего обнаружения, позволяют рентгенологам на раннем этапе различать злокачественные и доброкачественные новообразования в молочной железе.

Создание точных и эффективных классификаторов для больших баз данных является одной из основных задач научных исследований в области интеллектуального анализа данных и машинного обучения.

В настоящее время предложены много различных методов классификации, таких как деревья принятия решений, наивный - байесовский метод, и метод логистической регрессии, SVM, KNN и др.

Метод дерева принятия решений (j48) - обычно используется в интеллектуальном анализе данных для изучения данных и построения дерева и самих правил, которые будут использоваться для создания прогнозов.

Дерево решений - это классификатор, в виде древовидной структуры, где каждый узел является узлом листьев, указывающий значение целевого атрибута или класса примеров, или узел решение. Дерево принятия решений может использоваться для классификации объекта, путем перемещения начиная с корня дерева, пока не будет достигнут конечный узел, который обеспечивает классификацию экземпляра.

Наивный байесовский классификатор может быть как параметрическим, так и непараметрическим, в зависимости от того, каким методом восстанавливаются одномерные плотности. Основные преимущества наивного байесовского классификатора – простота реализации и низкие вычислительные затраты при обучении и классификации. В тех редких случаях, когда признаки действительно независимы, наивный байесовский классификатор оптимален. Основной его недостаток – относительно низкое качество классификации в большинстве реальных задач. Чаще всего он используется либо как примитивный эталон для сравнения различных моделей алгоритмов, либо как элементарный строительный блок в алгоритмических композициях [4].

Машина опорных векторов – является одной из наиболее популярных методологий обучения по прецедентам, предложенной В.Н. Вапником и известной в англоязычной литературе под названием SVM (Support Vector Machine). Это наиболее быстрый метод нахождения решающих функций. Метод сводится к решению задачи квадратичного программирования в выпуклой области, которая всегда имеет единственное решение. Не существует общего подхода к автоматическому выбору ядра в случае линейной неразделимости классов.

K-ближайших соседей – это метрический алгоритм классификации, основанный на оценивании сходства объектов. Классифицируемый объект относится к тому классу, которому принадлежат ближайшие к нему объекты обучающей выборки. Классификацию, проведенную данным алгоритмом, легко интерпретировать путём предъявления пользователю нескольких ближайших объектов. Поиск ближайшего соседа предполагает сравнение классифицируемого объекта со всеми объектами выборки, что требует линейного по длине выборки числа операций.

Метрики точности и полноты - Определим следующие величины:

TP - истинно положительные примеры - это количество случаев, правильно определенных как доброкачественные (true positives, TP):

TN - истинно отрицательные примеры - это количество случаев, правильно определенных как злокачественные (true negatives, TN):

FP - истинно положительные примеры - это количество случаев, неправильно определенных как доброкачественные (false positives, FP):

FN - ложные отрицательные примеры - это количество случаев неверно определенных как злокачественные (false negative, FN):

Точность представляет долю опухолей, которые были предсказаны как злокачественные от фактического числа злокачественных опухолей. Точность, специфичность и чувствительность вычисляется по следующим формулам:

$$(Точность)\,Accuracy = \frac{TP + TN}{TN + TP + FN + FP} \quad (1)$$

$$Специфичность(Specificity) = \frac{TN}{TN + FP} \quad (2)$$

$$Чувствительность(Sensitivity) = \frac{TP}{TP + FN} \quad (3)$$

Значение показателей точность, чувствительность и специфичность может быть определена с помощью матрицы смешивания, как показано в таблице 1. Матрица смешиваний описывает количество правильных и неправильных прогнозов, выявленных в системе классификации.

Таблица 1: Матрица смешиваний

	Фактически	
Фактический (actual)	Положительно	Отрицательно
Положительный	TP	FP
Отрицательный	FN	TN

Проводим оценку разработанной модели с использованием набора данных рака груди Висконсин для классификации поражения молочной железы.

Набор данных состоят из записей 699 пациентов. Среди них 458 или 65,5% имеют рак груди, для 241 или 34,5% результат неизвестен. Чтобы проверить результаты сравнения популярных шести методов интеллектуального анализа данных используется 10 - пересекающихся групп проверок.

При настоящем исследовании, данные разделены на 10 групп, где 1 группа используется для тестирования и 9 групп для обучения. Диагностические результаты записи каждого пациента из набора данных состоят из десяти переменных, которые кратко излагаются в таблице 2. Одна из 10 переменных является переменной, представляющей состояние диагностики пациента с или без рака груди (т.е. злокачественная или доброкачественная).

Таблица 2. Сведения об атрибутах.

№	Атрибут	Значение
1	Clump thickness (CT)	1-10
2	Uniformity of cell size (UCS)	1-10
3	Uniformity of cell shape (UCSh)	1-10
4	Marginal adhesion (MA)	1-10
5	Single epithelial cell size (SECS)	1-10
6	Bare nuclei (BN)	1-10
7	Bland chromatin (BC)	1-10
8	Normal nucleoli (NN)	1-10
9	Mitosis (M)	1-10
10	Class (C)	2/4

Таблица 2 сведения об атрибутах 2 для доброкачественных, 4 для злокачественных.

Степень точности классификации оценивается с точки зрения чувствительности и специфичности. Значения производительности методов (то есть точность, чувствительность, специфичность, коэффициент ошибок и время) определены на основе матрицы смешивания и показаны в таблице 3.

Таблица 3: Производительность обучения и тестирования данных

Алгоритм	Обучение данных (499)				Тестирование данных (200)			
	Acc	Senst	Spec	Err	Acc	Senst	Spec	Err
J48	95.59	0.96	0.949	4.41	92	0.942	0.873	8
NB	96.79	0.966	0.972	3.21	94.5	0.934	0.968	5.5
LR	96.79	0.978	0.949	3.21	92.5	0.956	0.857	7.5
SVM	97.59	0.981	0.966	2.41	94.5	0.956	0.921	5.5
KNN	95.19	0.978	0.904	4.81	94	0.949	0.921	6

Здесь Acc - точность, Senst - чувствительность, Spec - специфика, Err – коэффициент ошибок. Из приведенной выше таблицы видно, что метод SVM имеет наивысшую точность (97.59%) и наиболее низкий коэффициент ошибок (2,41%) как в обучении, так и в тестировании данных.

Выводы

В этой статье точность классификации методов оценивается на основании конкретных примеров. Важной задачей в области интеллектуального анализа данных и машинного обучения - построить точные и вычислительно-эффективные классификаторы для их применения в медицинской практике. Производительность метода SVM выше по сравнению с другими классификаторами. Следовательно, SVM показывает лучшие результаты для записей паци-

ента с заболеванием РМЖ. Поэтому классификатор SVM предлагается для диагностики РМЖ, так как позволяет получить результаты с высокой точностью, низким коэффициентом ошибок и высокой производительностью.

Литература

1. A. Endo, T. Shibata and H. Tanaka Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computing and Human Sciences, vol.13, pp.11-16. (2008)
2. Asuncion A. and D.J. Newman: "UCI Machine Learning Repository", Irvine, CA: University of California, School of Information and Computer Science, 2007. [Online]. Available: http://archive.ics.uci.edu/ml/datasets.html
3. О.Б. Рўзибоев, О.Қ. Хўжаев Исследование и программная реализация метода ближайших соседей// Химическая технология. Контроль и управление. – Ташкент, 2014. – №2. – С. 84-89. ISSN 1815-4840.
4. Hastie, T. The Elements of Statistical Learning / T. Hastie, R. Tibshirani, J. Friedman – Springer, 2001. – ISBN 0-387- 95284-5.

Е.Г. Супонев

АЛГОРИТМЫ СЖАТИЯ СИГНАЛОВ ЭЛЕКТРОКАРДИОГРАММ С ПОМОЩЬЮ ВСПЛЕСКОВ ДОБЕШИ

(Воронежский государственный университет)

Введение

В настоящее время компьютерные технологии широко применяются в исследованиях биологических систем. Одной из важных областей является электрокардиография (ЭКГ), изучающая активность сердечно-сосудистой системы человека.

Разработка эффективных алгоритмов сжатия сигналов ЭКГ обычно усложняется, во-первых, значительной вариабельностью и разнообразием признаков, во-вторых, наличием шумов от которых трудно избавиться на этапе регистрации сигнала, что обусловлено сложной природой явления [1,2]. Наличие этих факторов предполагает предварительную обработку сигнала перед компрессией. Поэтому актуальной задачей становится разработка универсальных алгоритмов и математических моделей, позволяющих повысить эффективность сжатия, при этом обеспечив максимальное качество результата.

1. Всплески с компактным носителем

Одной из тенденций настоящего времени в области цифровой обработки ЭКГ является применение теории всплесков [3]. Для осуществления сжатия удобно использовать всплески с компактным носителем, называемые всплесками Добеши.

Выражения для масштабирующей функции $\varphi(x)$ и всплеска $\psi(x)$ порядка $n/2$ (n – четное) выглядят следующим образом [4]