УДК 004.942

АЛЬЯНОВ А.Л., ВЛАСОВ В.В., ДЕМИДОВ А.В., МАМОШИН А.В.

К ВОПРОСУ О ПРИМЕНЕНИИ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ДИАГНОСТИРОВАНИЯ ФОРМ ПАНКРЕАТИТА

**Аннотация:**

В данной статье рассматривается подход для диагностирования вероятности развития разных форм деструктивного панкреатита, основанный на использовании методов машинного обучения. Для исследования процесса классификации пациентов были приняты популярные и часто применяемые на практике классификаторы:

* Метод опорных векторов;
* Метод ближайшего соседа;
* Ансамбль деревьев решений;
* Логистическая регрессия.

Для выбора лучшей прогностической модели для данной области медицинского исследования, имеющего свои индивидуальные особенности, в качестве критерия качества использовалась точность предсказания принадлежности пациента к определённой классификационной форме деструктивного панкреатита.

**Ключевые слова:**

Диагностика, панкреатит, классификация, Random Forest, SVM, k-Neighbors, logistic regression

**Введение**

Острый деструктивный панкреатит (ОДП) является одной из важнейших проблем неотложной хирургии органов брюшной полости. Это обусловлено, прежде всего, неуклонным ростом заболеваемости, увеличением частоты осложнений и высокой летальностью [1, 2, 3]. При ОДП летальность составляет 20-45% без дифференциации клинических форм, достигая при инфицированном панкреонекрозе 85%, а при фульминантном течении заболевания – 100% [1].

Анализ данных литературы позволяет признать, что клиника не располагает четкими критериями и методами дифференциального диагноза клинико-морфологических форм ОДП и его постнекротических осложнений. Ни один из существующих сегодня методов диагностики сам по себе не может быть признан абсолютно точным для раннего распознавания панкреонекроза и его осложнений. Следовательно, требуется комплексный подход в диагностике данного заболевания с учетом клинических и лабораторных данных, ультразвуковой томографии (УЗТ), лапароскопии, рентгенологических методов исследования, компьютерной томографии (КТ) и магнитно-резонансной томографии [1, 2, 3].

По нашему мнению, для ранней дифференциальной диагностики экссудативных образований, осложняющих течение ОДП, общепринятые методики обследования могут быть эффективно дополнены методом клиновидной дегидратации аспирата, полученного при проведении миниинвазивных вмешательств под контролем УЗТ и РТВ и миниинвазивной многочастотной биоимпедансометрией, способной вовремя миниинвазивной манипуляции в режиме реального времени определить состояние метаболизма зоны деструкции [4].

На основании имеющихся сведений в отечественной и зарубежной литературе актуальным представляется вопрос о своевременном выявлении инфицированного панкреонекроза и об аргументированном определении показаний к использованию контролируемых миниинвазивных хирургических методик на различных стадиях развития заболевания [5, 6].

Анализ литературы показал, что проблемой применения алгоритмов интеллектуальной обработки данных и машинного обучения занимаются многие ученые, которые делают акцент в основном на методы глубокого обучения с использованием искусственных нейронных сетей [7, 8, 9, 10].

**Основная идея**

Задача классификации пациентов представляет собой обучение модели на размеченных данных, которые представляют собой множество кортежей, где один атрибут отвечает за принадлежность объекта выборки к определенному классу, а все остальные атрибуты представляют собой характеристики объекта. После процедуры обучения модели на вход подаются данные похожего формата, который отличается только отсутствием атрибута принадлежности объекта к определенному классу. Задача прогностической модели на основе ранее учтенных размеченных данных дополнить неразмеченный кортеж новым атрибутом, что будет свидетельствовать о присвоении ему определенного класса.

В общем случае исходные данные для решений данной задачи классификации можно представить в виде двух матриц:

$X= \left(\begin{matrix}\begin{matrix}x\_{1, 1,g\_{1}}&x\_{1, 2, g\_{1}}\\x\_{2, 1,g\_{1}}&x\_{2,2,g\_{1}}\end{matrix}&\cdots &\begin{matrix}x\_{1,m,g\_{k}}\\x\_{2,m,g\_{k}}\end{matrix}\\\vdots &\ddots &\vdots \\\begin{matrix}x\_{n,1,g\_{1}}&x\_{n,2,g\_{1}}\end{matrix}&\cdots &x\_{n,m,g\_{k}}\end{matrix}\right) , Y= \left(\begin{matrix}y\_{1,1}&y\_{1,2}\\\begin{matrix}y\_{2,1}\\\vdots \end{matrix}&\begin{matrix}y\_{2,2}\\\vdots \end{matrix}\\y\_{n,1}&y\_{n,2}\end{matrix}\right)$,

где X - множество независимых переменных,

Y - множество зависимых переменных.

m - количество проводимых диагностических исследований над пациентом,

n - количество пациентов,

g – множество групп факторов,

k – количество групп факторов.

Говоря о множестве групп факторов g, мы подразумеваем множество $g\_{i} \in G=[анамнез, жалобы, объективно, ПОН, биохимический анализ крови, анализ мочи, С-реактивный, ОРОГ, ОРОБП, УЗИ, ЭГДС, ЭндоУЗИ, ПАБ БИ, ПАБ ЦИ, ПАБ АА, кристаллография, БИМ]$.

$Y\_{1}= (y\_{1,1}, y\_{2,1}..y\_{n,1})$ - множество значений принадлежности пациента к определенной форме деструктивного панкреатита,

$y\_{1,i}\in C\_{1}=[ОЖСК, Панкреонекроз, Панкреотогенный абсцесс, Киста]$, $Y\_{2}= (y\_{1,2}, y\_{2,2}..y\_{n,2})$ -  множество значений лежащих в интервале [0;1] свидетельствующих о инфицированной форме деструктивного панкреатита, $y\_{2,i}\in C\_{2}=[Инфицированный, Неинфицированный]$. Априорно известно, что в выборке пациентов всего два класса панкреонекроза: инфицированный и неинфицированный.

Данное исследование проводилось на данных, полученных в ходе врачебной практики. В этом наборе данных размер выборки представляет собой 324 наблюдения, каждый из которых характеризуется количеством атрибутов M = 290. Ознакомиться с данными можно в данном хранилище данных [11].

Процесс классификации пациентов начинался с предварительной обработки входных данных. В данной работе проводилось «обучение с учителем» и для обучения прогностической модели, т.е. ее конфигурирования, входной набор данных был разделен случайным образом на обучающую и тестовую выборки в соотношении 70% к 30% соответственно. Используя обучающую выборку, проводилось «обучение» классификационной модели, где в роли «учителя» выступали истинные значения из множества Y. После обучения для тестовой выборки, в которой истинные значения временно изъяты из рассмотрения, запускается процесс предсказания Ypred.

В результате получается множество истинных значений Y принадлежности пациента определенной клинико-морфологической форме из множества С и множества предсказанных значений Ypred.

$len\left(Y\_{ист}\right)= len\left(Y\_{пред}\right)=n$,

где len() – размерность множества.

Таким образом, данные можно представить в виде квадратной матрицы связности, которая устанавливает соответствие между истинными значениями Y и предсказанными Ypred на основе найденной в результате «обучения» функции зависимости. В общем виде матрица связности М выглядит следующим образом:

$$M=\left(\begin{matrix}\begin{matrix}m\_{1,1}&m\_{1,2}\\m\_{2,1}&m\_{2,2}\end{matrix}&\cdots &\begin{matrix}m\_{1,len(C)}\\m\_{2, len(C)}\end{matrix}\\\vdots &\ddots &\vdots \\\begin{matrix}m\_{len\left(C\right), 1}&m\_{len\left(C\right),2}\end{matrix}&\cdots &m\_{len\left(C\right), len(C)}\end{matrix}\right)$$

Исходя из точности классификации, с использованием полученной модели можно сделать вывод о значимости данного набора факторов, выраженного в количественных метриках. Перед расчётом метрик необходимо выбрать метод построения модели, который будет давать наиболее точное предсказание для данной области медицинского исследования. В качестве основной метрики оценки качества классификационной модели принято брать точность предсказания принадлежности пациента к определенной клинико-морфологической форме деструктивного панкреатита (Acc), показывающий какую долю дисперсии зависимой переменной объясняет выбранная прогностическая модель.

Наиболее оптимальным для оценки качества модели является метод кросс-валидации, т.к. обеспечивает репрезентативность обучающей выборки.

Отметим, что для построения модели были использованы следующие методы:

* [Метод опорных векторов](http://ru.wikipedia.org/wiki/%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD_%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD_%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD)
* [Метод ближайших соседей](http://ru.wikipedia.org/wiki/%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD_k_%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD_%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD%EF%BF%BD)
* [Ансамбли](http://ru.wikipedia.org/wiki/Random_forest) деревьев
* [Логистическая регрессия](http://ru.wikipedia.org/wiki/%D0%9B%D0%BE%D0%B3%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%B5%D1%81%D0%BA%D0%B0%D1%8F_%D1%80%D0%B5%D0%B3%D1%80%D0%B5%D1%81%D1%81%D0%B8%D1%8F)

**Результаты**

В ходе исследования прогностических моделей были получены следующие результаты, отображенные на рисунке 1. Точность модели определялась как отношение верно классифицированных пациентов к их общему числу.



Рисунок 1 – Качество прогностических моделей

Как видно из вышеприведенного рисунка, из применяемых методов наилучшие результаты показала модель, построенная на основе метода Random Forest. Был построен двумерный график ROC-кривой, где по оси абсцисс откладывается TPR, а по оси ординат FPR. Ключевым параметром для оценки точности прогностической модели является AUC (площадь под кривой).



Рисунок 2 - ROC-кривая

Так как количество классов равно количеству форм панкреонекроза (n = 4), то был применен мультиклассовый ROC-анализ, который сводится к бинарной классификации путем упрощения, например, «класс 1»/«все остальные классы» [4, 5].

Степень влияния каждого фактора при классификации можно выделить, используя веса узлов дерева. Влияние факторов на результат классификации приведено ниже. Отметим некоторые самые влиятельные факторы:

 0.072499 - PAB - цитологическое исследование аспирата липофаги.

 0.043055 - Биоимпедансометрия - Измерение комплексного сопротивления тела человека при прохождении через него разночастотного переменного электрического тока. Частота 100 КГц.

 0.042055 - PAB цитологическое исследование аспирата липофаги кисты

 0.037939 - BIM Биоимпедансометрия. Частота 10 КГЦ.

 0.035442 - Кристаллография

 0.033691 - UZI УЗИ

 0.026910 - BIM Биоимпедансометрия дистально, частота 1 КГц.

 0.024123 - BIM Биоимпедансометрия проксимально, частота 100 КГц.

 0.023968 - BIM Биоимпедансометрия проксимально, частота 10 КГц.

 0.023042 - Кристаллография, соотношения площади белковой зоны к общей площади фации.

Кроме того, в результате исследования, были получены модели по отдельным группам диагностических исследований для выяснения возможности классификации деструктивной формы панкреатита для упрощенной диагностики заболевания. Особое внимание уделяется неинвазивным методам, таким как биоимпедансометрия, кристаллография. Выбор методов обусловлен клинической практикой автора (Мамошин А.В.), позволившей наблюдать вклад данных методов в общую задачу классификации.

**Заключение**

В качестве прогностической модели в будущем видится перспективно использовать глубокое обучение посредством многослойной нейронной сети. Данный метод обладает высокой степенью точности и способен выявлять скрытые взаимосвязи между признаками, что может дать положительный результат в медицинской практике диагностирования панкреонекроза. Однако во избежание проблемы недообучения сети необходимо на этапе обучения модели иметь более репрезентативную и крупную по размеру выборку. Вследствие того, что некоторые признаки коррелируют между собой, тем самым негативно влияя на точность предсказания, необходимо использовать методы понижения размерности признакового пространства, например с использованием метода главных компонент (PCA).

**Список использованной литературы:**

1. Allwein, E.L., Schapire, R.E. and Singer, Y. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research. Pages 113–141.

2. Fawcett, Tom (2006). An introduction to ROC analysis. *Pattern Recognition Letters.* Volume 27, Issue 8., Pages 861-874.

3. Breiman, Leo (2001). Random Forests. *Machine Learning 45*. Pages: 5–32. DOI:10.1023/A:1010933404324

4. Zweig M.H., Campbell G. (1993). ROC Plots: A Fundamental Evaluation Tool in Clinical Medicine. *Clinical Chemistry*. Vol. 39, No. .

5. Davis J., Goadrich M. (2006). The Relationship Between Precision-Recall and ROC Curves // Proc. Of 23 International Conference on Machine Learning, Pittsburgh, PA, 2006

6. Mark J. L. Orr (1996). Introduction to Radial Basis Function Networks Centre for Cognitive Science, Centre for Cognitive Science, University of Edinburgh.

7. Yang Fei,Jian Hu,Kun Gao,Jianfeng Tu,Wei-qin Li,Wei Wang (2017). Predicting risk for portal vein thrombosis in acute pancreatitis patients: A comparison of radical basis function artificial neural network and logistic regression models. Journal of Critical Care. Elsevier.

8. Jianwei Zhu, Lei Wang, Yining Chu, Xiaojia Hou, Ling Xing, Fanyang Kong, Yinghuo Zhou, Yuanyuan Wang, Zhendong Jin, Zhaoshen Li (2015). A new descriptor for computer-aided diagnosis of EUS imaging to distinguish autoimmune pancreatitis from chronic pancreatitis Gastrointestinal Endoscopy, Volume 82, Issue 5, Pages 831-836.

9. Bodil Andersson,Roland Andersson,Mattias Ohlsson,Johan Nilsson (2011).
Prediction of Severe Acute Pancreatitis at Admission to Hospital Using Artificial Neural Networks. Pancreatology. Elsevier.

10. Aravind Sugumar,Metab Alkubeyyer,Bradley Erickson,Santhi Swaroop Vege (2011). Can Machine Learning Predict the Severity of Acute Pancreatitis. Gastroenterology. Elsevier.

11. «Лаборатория биоинформатики и интеллектуальных методов обработки биомедицинских данных» (2017), Science Data Storage“Pancreatitis”, Mendeley Data, v1 http://dx.doi.org/10.17632/gzxpmy8h8z.1 DOI is reserved but not active.