О задаче классификации для диагностики тяжести острого панкреатита (в системе поддержки принятия врачебных решений)

Мангалова Е.С[1]., Строев А.В.[2], Чубарова О.В.[3]

1 - ООО «Ар Ди Сайнс, Красноярск 2 – ФГБОУ ВО Красноярский государственный медицинский университет имени проф. В. Ф. Войно-Ясенецкого Министерства здравоохранения РФ, 3 – ФГБОУ ВО «Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева» », Красноярск

Аннотация

Целью исследования является улучшение результатов лечения пациентов с острым панкреатитом путем совершенствования объективизации степени тяжести. Для этого использовались методы интеллектуального анализа данных, в частности решалась задача классификации. Исходный объем данных составил 130 историй болезни. Проводилась предварительная обработка «сырых» данных на предмет заполнения пропусков в данных. С этой целью использовались разные подходы в зависимости от возможности их применения: восстановление медианой, линейная регрессия, логика.

Острый панкреатит является актуальной проблемой неотложной абдоминальной хирургии. В течение нескольких лет в ряде регионов России острый панкреатит занимает лидирующие позиции в структуре острой хирургической патологии, в свою очередь смертность при деструктивных формах острого панкреатита стабильно растет. Объективное оценка тяжести заболевания имеет чрезвычайно важное значение в лечении больных, так как определяет возможность назначить своевременное корректное лечение. Существующие прогностические системы (известно более 20 прогностических систем) при остром панкреатите громоздки, трудоемки, не всегда точны. Большинство из них включают сложные критерии, которые невозможно использовать в условиях городских и районных стационаров, учитывая тот факт, что большая часть больных поступает в вечернее или ночное время, при этом исследователи рекомендуют оценить прогноз течения заболевания от начала приступа острого панкреатита на протяжении 24 часов. Только врач с обширной практикой и длительным опытом способен в первые сутки определить степень тяжести, которая в дальнейшем подтверждается. Применение в клинической практике современных компьютеров позволяет расширить возможности решения задач прогнозирования течения и исхода острого панкреатита в силу доступности анализа накопленных данных по составу и числу групп параметров заболевания.

Таким образом, возникает задача классификации вновь прибывшего пациента, по данным о состоянии пациента только в первые сутки поступления в больницу. Тогда задача классификации может быть сформулирована следующим способом: имеется множество историй болезни пациентов с диагнозом острый панкреатит с выставленными степенями тяжести заболевания (классами), то есть задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество будем называть выборкой. Требуется построить алгоритм, способный классифицировать вновь поступившего пациента с диагнозом острый панкреатит, то есть определить степень тяжести заболевания.

Для решения задачи классификации в первую очередь необходимо выявить значимые, с позиции решения задачи, показатели (признаки) состояния пациента. Присутствие в данных неинформативных (не значимых) признаков приводит к снижению точности решений, увеличению вычислительных процедур. Отбор признаков позволяет по минимально возможному набору показателей выявить степень тяжести, а значит назначить своевременно корректное лечение. Задача классификации решалась при помощи нескольких алгоритмов: ridge regression, SVM, нейронные сети.

Имеется множество историй болезни пациентов с диагнозом острый панкреатит с выставленными степенями тяжести заболевания (классами), то есть задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество будем называть выборкой. Требуется построить алгоритм, способный классифицировать вновь поступившего пациента с диагнозом острый панкреатит, то есть определить степень тяжести заболевания.

Для решения задачи классификации в первую очередь необходимо выявить значимые, с позиции решения задачи, показатели (признаки) состояния пациента. Присутствие в данных неинформативных (не значимых) признаков приводит к снижению точности решений, увеличению вычислительных процедур. Отбор признаков позволит по минимально возможному набору показателей выявить степень тяжести, а значит назначить своевременно корректное лечение.

В исходной выборке содержалось 27 показателей состояния пациента. Все показатели можно разделить на группы: 11 показателей расширенного анализа крови (РАК), 4 показателя анализа мочи, 8 показателей УЗИ, температура, характер перистальтики, вздутие, и оценка тяжести, данная экспертом. 7 показателей являлись категориальными переменными (принимали значение 0, 0.5 либо 1). Основная часть категориальных переменных имела значения 0 и 1, то есть наличие или отсутствие признака, а один – вздутие, принимал значения 0, 0.5 и 1, где 0 – отсутствие признака (живот не вздут), 1 – признак присутствует (живот вздут) и 0.5 – характеристика переменной типа живот умеренно вздут, подвздут и т.п. Кроме того, данные содержали пропуски.

Для работы с данными в первую очередь необходимо было по возможности восстановить пропуски. Заполнение пропусков показателей, имеющих числовые значения, проводилось на основе линейной регрессии в случае подтверждения гипотезы линейности. Были восстановлены значения показателя гемоглобин на основе значения гематокрит, позднее показатель гематокрит был удален из выборочных данных, как менее заполненный исходными данными. Остальные пропуски в числовых данных заполнялись расчетным значением медианы показателя. Восстанавливая данные, рассматривались гистограммы распределений значений признака, если наблюдались классовые смещения, то заполнение медианой проводилось внутри класса.

При восстановлении данных характеризующих размеры поджелудочной железы (голова, тело и хвост) использовалась линейная регрессия. Данный метод заключается в том, что пропущенные значения заполняются с помощью модели линейной регрессии, построенной на известных значениях набора данных. Таким образом, было восстановлено 34 пропуска (26% выборки). В большинстве случаев восстанавливалось значение величины хвоста по известным значениям головы и тела.

Заполнение пропусков в категориальных данных осуществлялось в соответствие со следующим правилом: отсутствие записи в истории болезни информации в разделе УЗИ (КТ), означает отсутствие отягощающего признака, иначе врач/диагност это указывают в обязательном порядке. Поэтому логично заполнить пустое значение 0, интерпретируя его как «Отсутствует».

В исходную выборку входят переменные, измеренные в разных шкалах, некоторые из них являются категориальными, как следствие, размах значений некоторых также существенно разнится. Для приведения всех переменных к одинаковым единицам измерения служит процедура нормализации. Помимо того, что многие переменные выражаются в разных единицах, некоторые из них имеют также явно выраженные асимметричные распределения. Часто решить эту проблему позволяют такие простые преобразования исходных значений как квантильное преобразование (метод обратного преобразования).

В исходных данных экспертами были расставлены метки принадлежности точек (пациентов) к классам. Всего определили 3 класса (1 степень тяжести – легкая форма, 3

степень тяжести – тяжелая форма и 2 степень тяжести). Для визуализации исходных данных был применен алгоритм t-SNE.

При решении практических задач классификации и восстановлении неизвестной стохастической зависимости с помощью регрессии предлагается использовать порожденные признаки, полученные с помощью измеряемых исходных признаков. Это влечет существенное повышение размерности признакового пространства и, как следствие, необходимость использования алгоритмов выбора признаков.

Одним из методов понижения размерности является гребневая регрессия (ridge regression). Гребневую регрессию используют, если имеет место:

переизбыточность данных;

коррелированность независимых переменных;

сильные различия собственных значений характеристического уравнения или близость к нулю некоторых из них.

В качестве критерия качества использовалась оценка площади под кривой AUC(Area Under Curve). Значение этой оценки может меняться от 0 до 1, но, как правило, говорят об изменениях от 0.5 («бесполезный» классификатор) до 1 («идеальная» модель).

Исходная выборка (объем данных 130) была разделена на обучающую (100 элементов) и тестовую (30 элементов), всего количество комбинаций составило 10000.

Настройка регуляризующего параметра осуществлялась перебором, параметр принимал значения из диапазона [0:1000]. С каждым значением регуляризующего параметра строилась гребневая регрессия. Если с увеличением значения параметра точность увеличивается, следовательно, продолжаем брать большие значения, в варианте незначительного увеличения точности возвращаем предыдущее значение.

Далее с целью понижения размерности признакового пространства применялась гребневая регрессия в комбинации с методом оптимального прореживания. Оптимальное прореживание — метод упрощения структуры регрессионной модели. Основная идея прореживания: элементы модели, которые оказывают малое влияние на ошибку, можно исключить из модели без значительного ухудшения качества.

Из структуры гребневой регрессии удалялись по очереди признаки, коэффициенты которых были наименьшими, если при этом точность улучшалась или не значительно ухудшалась, признак удалялся навсегда, в противном случае – возвращался в регрессию.

В результате всех, описанных действий, удалось сократить пространство признаков до 12. Задачу классификации решали на основании, полученных регрессионных зависимостей. Процент ошибки классификации составил менее 6%.

Список использованной литературы:

1. Галимзянов Ф.В. Первичная диагностика инфицированного панкреонекроза // Хирургия им. Н.И. Пирогова. – 2006. – № 6. – С. 8–10.

2. Стрижов, В.В. Методы выбора регрессионных моделей / В.В. Стрижов, Е.А. Крымова - М.: Вычислительный центр РАН, 2010 – 60 с.

3. Диагностика и лечение острого панкреатита (Российские клинические рекомендации)// Российское общество хирургов, Ассоциация гепатопанкреатобилиарных хирургов стран СНГ, Российское общество скорой медицинской помощи, принято 30 октября 2014г.

4. van der Maaten, L Visualizing Data using t-SNE /L. van der Maaten, G.Hinton – Jurnal of Machine Learning research 9 (2008), p. 2579-2605

5. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Книга 2. М.: Финансы и статистика, 1986. — 351 с.

6. Вадзинский Р.Н. Справочник по вероятностным распределениям. - СПб.: Наука, 2001, 295