УДК 004.89

ИССЛЕДОВАНИЕ МЕТОДА ОПОРНЫХ ВЕКТОРОВ В ЗАДАЧЕ ИДЕНТИФИКАЦИИ КЛЕТОК ТРОМБОЦИТОВ

Елагина Е. А. (Национальный исследовательский университет ИТМО)
Научный руководитель – к.т.н, доцент Маргун А.А.
(Национальный исследовательский университет ИТМО)

Существует ряд заболеваний, которые можно диагностировать на ранних стадиях с помощью анализа крови, в частности, при оценке свертывающей способности крови в первую очередь определяется уровень тромбоцитов. В настоящее время выделение и подсчет количества клеток крови при лабораторном анализе биологического материала медицинским работником является трудозатратным процессом. Развитие области медицинской визуализации становится важной задачей вследствие растущей потребности в автоматизированной, быстрой и эффективной диагностике. В представленной работе решение данной проблемы предлагается на основе систем искусственного интеллекта, в частности, с использованием метода опорных векторов (SVM), представляющего собой усовершенствованную технику на основе ядра и используется для классификации данных.

Введение. В данной работе рассматривается приложение метода опорных векторов к процессу автоматизации идентификации и подсчета количества клеток тромбоцитов. Визуализация помогает врачам анализировать изображения, выявлять аномалии внутренних структур, что особенно важно при исследовании мазков периферической крови. Подсчет количества тромбоцитов важен, поскольку они выполняют две основных функции: формирование тромбоцитного агрегата – первичной пробки, закрывающей место повреждения сосуда – и предоставление своей поверхности для ускорения ключевых реакций плазменного свёртывания. Относительно недавно установлено, что тромбоциты также играют важнейшую роль в заживлении и регенерации повреждённых тканей, выделяя из себя в повреждённые ткани факторы роста, которые стимулируют деление и рост клеток. Факторы роста представляют собой полипептидные молекулы различного строения и назначения.

Основная часть. В данном исследовании решается задача разработки нейросети для идентификации и подсчета тромбоцитов. Существует множество методов решения данной задачи, на основе проведенного аналитического обзора было принято решение использовать метод опорных векторов (support vector machines, SVM). SVM – очень мощный и гибкий класс алгоритмов обучения с учителем как для классификации, так и регрессии. Возможности данного метода расширяются при его комбинации с ядрами (kernels), что позволяет проецировать данные в пространство с большей размерностью, определяемое полиномиальным и Гауссовыми базисными функциями, благодаря чему появляется возможность аппроксимировать нелинейные зависимости с помощью линейного классификатора. Очевидно, что есть данные, не допускающие линейного разделения, но их можно спроецировать в пространство более высокой размерности, а следовательно, будет достаточно линейного разделителя. Потенциальная проблема, возникающая при использовании указанной методики, заключается в том, что при проецировании N точек на N измерений могут потребоваться колоссальные объемы вычислений. Однако, благодаря процедуре kernel trick обучение на преобразованных с помощью ядра данных можно провести неявно, то есть даже без построения полного N-мерного представления ядерной проекции. Эта процедура является частью SVM и одним из больших преимуществ метода.
Реальные наборы данных часто бывают зашумлены и неоднородны, в них могут отсутствовать признаки. Одним из очень интересных приложений машинного обучения является анализ изображений с использованием пиксельных признаков для классификации. На практике анализируемые данные очень редко оказываются достаточно однородными, и

простых пикселей будет недостаточно. Это привело к появлению методик выделения признаков, например, с помощью гистограммы направленных градиентов (HOG), которая преобразует пиксели изображения в векторное представления, чувствительное к несущим информацию признакам изображения.
Цель извлечения признаков – уменьшение исходного набора данных путем измерения определенных свойств или функций, которые отличают один входной шаблон от другого. Извлеченные признаки становятся входными данными для классификатора, который будет считаться соответствующими свойствами изображения в пространстве признаков. Если доступно только несколько обучающих образов, каждая из структур может иметь несколько обучающих примеров, в которых фактически присутствует тромбоцит (PLT). После первоначального выбора потенциально хороших функций наиболее целесообразно выбрать небольшой набор «хороших» функций PLT. Алгоритм автоматического выбора функций может использоваться для дальнейшего сужения до набора, состоящего из наиболее важных из них.
Для реализации метода был выбран язык программирования Python. С использованием метода опорных векторов (SVM), дополненного гистограммой направленных градиентов (HOG) для определения признаков, была построена и обучена нейронная сеть. Работоспособность метода протестирована на имеющемся тестовом наборе данных.

Выводы. Классификация методом опорных векторов (SVM), по данным более ранних исследований, достигала точности 99,8%. Следовательно, можно сделать вывод о том, что SVM является мощным методом классификации по ряду причин:

1. Зависимость метода от относительно небольшого количества опорных векторов означает компактность модели и небольшого объема используемой оперативной памяти.
2. Фаза предсказания после обучения модели занимает очень мало времени.
3. Этот метод хорошо походит для многомерных данных, в том числе, с количеством измерений большим, чем количество выборок, что является непростым условием работы для других алгоритмов.
4. Интеграция с ядерными методами делает метод универсальным, обеспечивает приспособляемость к множеству типов данных.

Следует заметить, что у метода имеются и недостатки:

1. При значительном количестве обучающих выборок вычислительные затраты могут оказаться непомерно высокими.
2. Результаты зависят от выбора параметра размытия *C*. Его необходимо выбирать с помощью перекрестной проверки, которая также может потребовать значительных вычислительных затрат при росте размеров наборов данных.
3. У получаемых результатов отсутствует непосредственная вероятностная интерпретация. Ее можно получить путем внутренней перекрестной проверки, но это также потребует больших вычислительных затрат.