ПОДГОТОВКА ВЕКТОРА ВХОДНЫХ ДАННЫХ КАК ВАЖНЫЙ АСПЕКТ РАЗРАБОТКИ СИСТЕМ ОБРАБОТКИ ИНФОРМАЦИИ

Коноваленко С.Н., Мороз Б.И.

Академия таможенной службы Украины, г. Днепропетровск

В работе выделена проблематика выбора методов подготовки сырых данных для последующей их обработки в системах интеллектуального анализа информации. Любая предметная область содержит в себе массу разнотипных идентификационных характеристик. В качестве первичного источника данных могут выступать хранилища и базы данных коммерческих и государственных организаций, представляемые документы, сеть интернет, т.е. любая информация, которая может пригодиться для принятия решения.

Большинство алгоритмов неспособны напрямую работать со всеми типами и видами данных. Поэтому подготовка или трансформация входных наборов данных является неотъемлемой частью проектирования систем анализа информации. Не сделав этого, мы ухудшаем качество работы системы анализа информации (распознавание образов, классификация и т.д.), а в некоторых случаях это приведёт даже к невозможности адекватно воспринимать входной вектор данных. Препроцессинг – это процедура подготовки данных к анализу, в процессе которого они приводятся в соответствие с требованиями, определяемыми спецификой решаемой задачи (предметной областью) и используемой моделью обработки (анализа) полученной информации. Как правило, предобработка данных включает два направления:

1. Очистка и оптимизация.
2. Трансформация, нормализация.

В работе рассмотрены методы преобразования непрерывных и дискретных видов данных в пригодные для анализа вектора и множества. На примере предметной области «Информация таможенного контроля» было показано, как формируется обучающее множество для системы распознавания рисков нарушения таможенного законодательства на основе нейронной сети типа многослойный персептрон. Таким образом, входной вектор был преобразован к единому формату и диапазону – [0…1], который подходит для активационной функции. Далее он может быть подан на вход используемой нейронной сети для обучения и классификации. Также были рассмотрены методы формирования псевдографических образов из входной последовательности данных изначально не графического происхождения. Далее эти графические образы подаются на вход нейронной сети для обучения и распознавания.

В результате проделанной работы были рассмотрены методы и средства подготовки входного вектора данных для системы идентификации рисков нарушения таможенного законодательства. В результате чего улучшилась репрезентативность и качество обучающей выборки, как для нейроклассификатора, так и для проектировщика системы анализа информации.