Секция «Математика и механика»

Использование потенциальных функций для обобщения байесовского подхода к построению одноклассового классификатора в задаче фильтрации нежелательной почты

Бурмистров Михаил Олегович
Студент
Московский физико-технический институт, факультет управления и прикладной математики, Долгопрудный, Россия
E-mail: i.like.spam@ya.ru

В современном мире электронная почта широко распространена и используется во многих сферах человеческой жизни. В силу своей высокой открытости этот канал обмен сообщениями стал активно использоваться мошенниками и злоумышленниками, что поставило задачу автоматической фильтрации спама. Эта задача решается различными методами, использующие как эвристические [2,4], так и вероятностные постановки. При этом отдельно стоит задача составления «хорошей» обучающей выборки, поскольку классы электронных писем «спам»/«не спам» существенно разнородны. Так, письма, представляющие интерес для пользователя, обладают

- меньшей доступностью,
- высокой разнородностью,
- большим числом шаблонных писем (разнообразные уведомления от сервисов).

По этим причинам представляется разумным построить классификатор, использующий для обучения лишь объекты одного из классов. В данной работе этот классификатор обучается на письмах, являющимися спамом. В предыдущей работе [1] был предложен Байесовский подход к решению задачи одноклассовой классификации. Решающее правило при этом оказывается основано на определении принадлежности точки в признаковом пространстве, соответствующей рассматриваемому объекту, гиперсфере, радиус и центр которой определяются на этапе обучения модели. Этот результат совпадает с эвристической постановкой задачи, предложенной в [3]. Однако, использование гиперсферы в качестве разделяющей поверхности оказывается слишком жестким условием, и качество классификации можно повысить, введя в модель потенциальные функции. При этом классификатор становится способен строить более богатое семейство разделяющих поверхностей, что позволяет повысить обобщающую способность. В работе приведено соответствующее обобщение модели и проведены вычислительные эксперименты на модельных и реальных данных, показывающие эффективность использования такого подхода.

Литература

1. Бурмистров М. О., Сандуляну Л. Н. Вероятностная модель одноклассовой классификации // Машинное обучение и анализ данных, М., 2012. Т. 1, №4, стр. 420–427.
2. R. Islam, U. Chowdhury Spam filtering using ML algorithms // Universitetets Okonomiske Institute, IADIS International Conference on WWW/Internet, 2007.

3. D. Tax One-class classification; Concept-learning in the absence of counterexamples, Ph.D thesis, 2001.

4. J. Sun, Q. Zhang, Z. Yuan, W. Huang, X. Yan, J. Dong Research of Spam Filtering system based on LSA and SHA // Advances in neural networks — ISNN 2008, 2008.