УДК 004.032.26 + 004.82

И.С. Мисуно, Д.А. Рачковский, С.В. Слипченко

РАСПРЕДЕЛЕННОЕ ПРЕДСТАВЛЕНИЕ ДАННЫХ В ЗАДАЧАХ КЛАССИФИКАЦИИ

Введение

Задача классификации заключается в отнесении образцов входных данных к одному или нескольким классам из заранее определенного набора [1]. В индуктивном подходе классификация осуществляется на основе обучающей выборки, которая содержит примеры образцов данных с заранее назначенными метками классов. Часто входные данные представлены в виде вектора числовых величин. Элементы вектора – действительные числа (например, результаты измерений некоторых характеристик объекта или их функции), или бинарные величины (индикаторы наличия некоторых признаков во входных данных).

Такая входная векторная информация зачастую не содержит в явном виде релевантной для классификации информации, поэтому полезной является ее трансформация. Нами разработаны методы трансформации входной информации (числовой [2], текстовой [3], зрительной [4]) в бинарные распределенные представления. Для классификации этих представлений можно использовать линейные классификаторы – как *SVM* [5], так и более вычислительно эффективные и естественно работающие с многими классами персептроноподобные [4, 6]. Задачей данной работы является исследование эффективности предложенных методов распределенного представления информации, а также модификаций методов классификации для искусственных и естественных данных разных модальностей.

Классификация числовой векторной информации.

Для исследования методов распределенного представления и классификации числовой информации были выбраны следующие известные тестовые задачи: Леонарда-Крамера *LK*, исключающее ИЛИ, двойная спираль [6], данные, генерируемые *DataGen* [6], а также данные базы *Elena* [7]. Размерность A векторов данных составляла от 2 до 36, число классов C – от 2 до 11, число образцов в обучающей и тестирующей выборках – от 75 до 3218.

Все выбранные задачи имеют существенно нелинейные области классов. Поэтому в качестве нелинейного преобразования использовались методы кодирования числовых векторов рецептивными полями *RSC* и *Prager* [2], которые выделяют бинарные признаки – индикаторы попадания входного A-мерного вектора в s-мерные ($s<A$) гиперпрямоугольные поля со случайным расположением и размером. Пусть общее число признаков (полей) N, из них среднее число единичных признаков $M=Np$. Плотность кода p управляется параметром – средним размером поля. С p связана разрешающая способность кодирования (средний размер *cell* элементарной ячейки – области входного пространства, где код не изменяется), а также распределение $P(s)$ размерности полей s ($s=0,\ldots,S$), где $S<A$ – максимальная заданная размерность полей. При фиксированной p увеличение N приводит к уменьшению *cell*. В [2] получены аналитические выражения для вычисления этих параметров, а также для зависимости перекрытия кодов от координат точек входного пространства. Последняя зависимость может рассматриваться в качестве ядра [5]. Использование ядра эквивалентно использованию кодов с бесконечно большим N и $cell=0$. Поэтому здесь p оказывает влияние только на форму характеристики перекрытия, которая определяется пропорцией полей разной размерности $P(s)$.

Для исследования влияния данных параметров на качество классификации использовалась следующая схема экспериментов. Векторы входных данных тестовых задач преобразовывались в коды *RSC* и *Prager* с разными параметрами. Эти коды использовались в качестве входной информации для линейных классификаторов: коды обучающих выборок для обучения, коды текстирующих выборок – для тестирования. Показателем качества классификации служил процент ошибок при тестировании.

В качестве линейных классификаторов использовались *SVM* [5], а также варианты разрабатываемых нами персептроноподобных классификаторов [4]. Кроме того, классификация осуществлялась с использованием ядерного *SVM* – на полученных нами ядрах (*RSC* и *Prager*), а также на стандартных ядрах (гауссово, полиномиальное). Известно, что *SVM* не обучается в ходе процесса, требует решения вычислительно сложной задачи нелинейного программирования, а также проводит оптимальную поверхность только для двух классов. В

 ISSN 1562-9945

данной работе для преодоления недостатков *SVM* применен разработанный нами персептрон с защитной полосой и правилом одновременного обучения на несколько классов. В нем выходной сигнал нейронов, соответствующих классам, определяется как $y_c = \Sigma_i\ x_i w_{ic}$, где w_{ic} – веса обучаемых связей, x_i – значение i-го элемента входного вектора. Для нейрона правильного класса $y_{c\text{-}true} = y_{c\text{-}true}(1-T)$, $0<T<1$ – величина "защитной полосы". Результатом классификации является индекс (номер) c^* нейрона с максимальной активацией: $c^* = \mathrm{argmax}_c\ y_c$. При ошибке ($c^* \neq c_{true}$) связи модифицируются следующим образом: $w_{ic} = w_{ic} + \Delta w$ для $c=c_{true}$ и $w_{ic} = w_{ic} - f(\Delta w)$ для c: $y_c > y_{c\text{-}true}$, где c_{true} – индекс правильного класса. Например, $f(\Delta w) = \Delta w/|c|$.

Результаты экспериментального исследования. На рис. 1 приведены результаты для задачи Леонарда-Крамера – зависимость процента ошибок классификации % *errors*, размера элементарной ячейки *cell* и средней размерности полей $E\{s\}$ от плотности кода p. Для кодирования *Prager* и *RSC*, для *SVM* и персептрона с защитной полосой (T=0.75) приведены усредненные результаты по 10 реализациям кодов при N=100. Также приведены результаты для *SVM* с ядрами (*Kernel*). Для всех случаев (а также для больших значений N, не показанных на рисунке) минимум ошибок достигается вблизи p=0,25, что соответствует минимальному значению *cell* и $E\{s\}$=2.

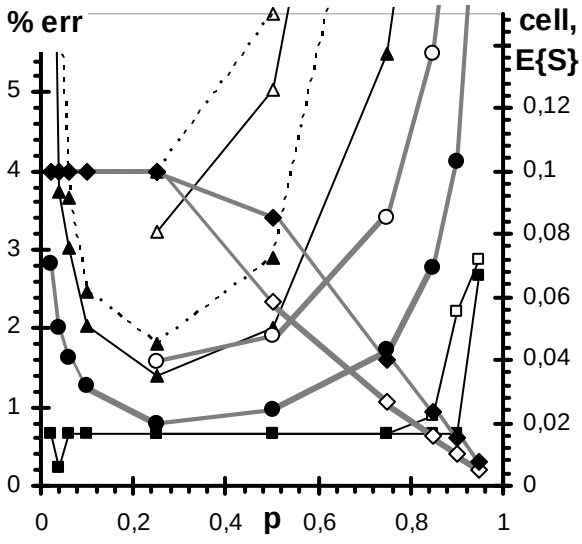

Рис. 1

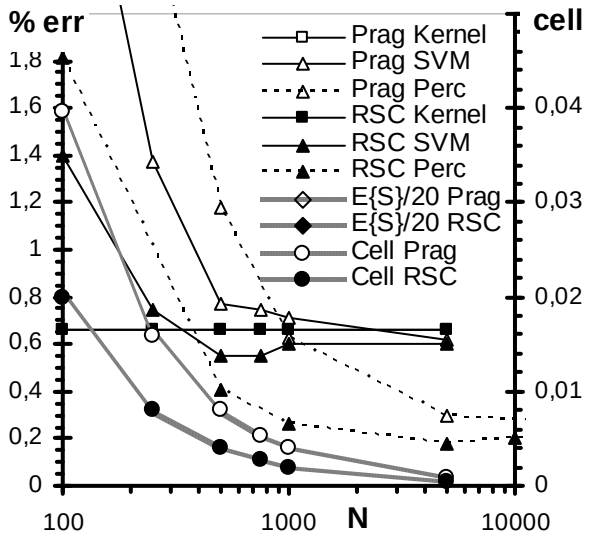

Рис. 2

Результаты рис. 2 показывают зависимость % *errors* и *cell* от N при p=0,25. Приведены усредненные результаты по 10 реализациям кодов. Уже при N=500 результат для *SVM* близок к результату для

ядра. Для персептрона с полосой (T=0.75) ошибка при N>(300–1000) ниже, чем для *SVM*. Время обучения для персептрона меньше, чем для *SVM* в 20 раз, а тестирования – в 100 и более раз.

Результаты экспериментов для данных *DataGen* приведены на рис. 3 при A=4, S=3, C=4, R=4 (параметр, регулирующий сложность областей классов [6]), число образцов на класс равно 100. Усреднение проводилось по 5 реализациям точек *DataGen* и 5 реализациям кодов. Для таких параметров минимальное значение *cell* соответствует p~0.3 и близко к нему в интервале p=0.125...0.5. Минимум ошибки для *SVM* и для персептрона с полосой также достигается в этом интервале. При этом для N=100 он смещен в сторону больших значений p (что обеспечивает большую стабильность M). При N=1000 минимум смещается в сторону меньших p, где обеспечивается большая реальная размерность рецептивных полей, M остается достаточно большим, а *cell* – малым. Для персептрона с полосой (T=0.75) достигаемый минимум ошибки соответствует уровню *SVM*, однако при отклонении от минимума ошибка для *SVM* увеличивается медленнее. Отметим, что кодирование с S=2 и S=5 дает сравнимые результаты классификации. Время обучения для персептрона в 20 раз, а тестирования примерно в 500 раз меньше, чем для *SVM*.

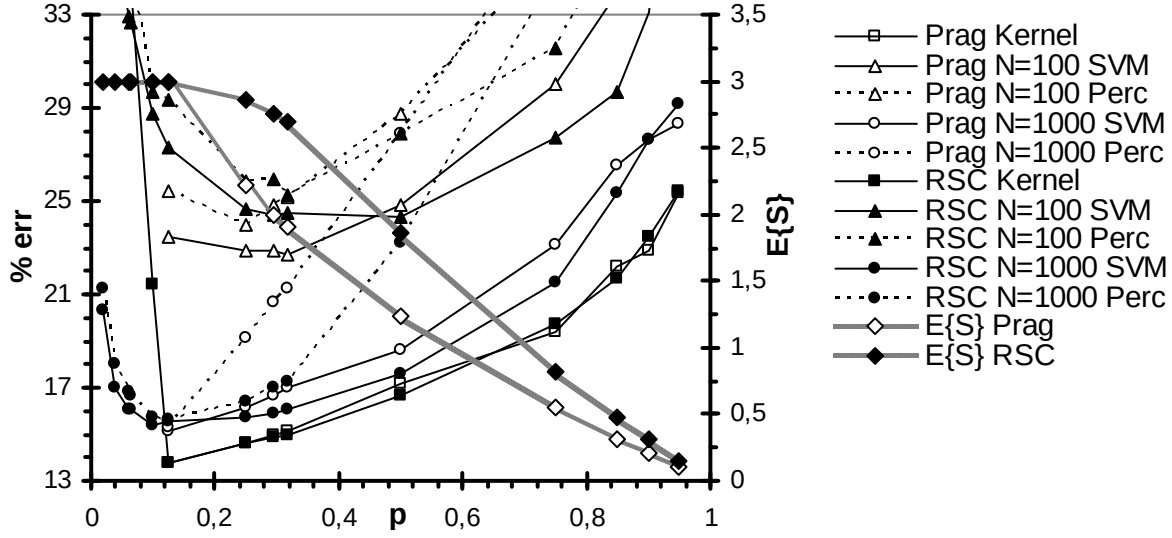

Рис. 3

Для искусственных данных базы *Elena* использовались параметры кодирования N=1000, A=S=2, p=0.25; для естественных (*Iris*, *Phoneme*, *Satimage*, *Texture*) – N=10000, S=2, 5(4), p=0.1 или 0.25. В таблице 1 приведены результаты исследования – процент ошибок

 ISSN 1562-9945

классификации. Для *SVM* и персептрона результаты получены усреднением по 10 реализациям кодов *RSC* и *Prager*. Там же приведены лучшие результаты известных методов *kNN*, *MLP*, *IRVQ* (см. [7]). Сравнение результатов показывает, что с помощью кодирования *RSC* и *Prager* получен лучший результат на *Concentric*, *Phoneme*, *Texture* и второй результат для *Satimage* и *Gaussian 7D*. Для остальных задач результаты незначительно уступают лучшим методам: *kNN* с небольшим отрывом лидирует на естественных данных *Satimage* и *Iris*, *MLP* и *IRVQ* - на искусственных *Gaussian* и *Clouds*. Время обучения для персептрона в среднем в несколько раз меньше, чем для *SVM*, а тестирования – в десятки раз меньше.

Таблица 1

Наборы данных	RSC SVM	RSC ядро	RSC перс.	Prager SVM	Prager ядро	kNN	MLP	IRVQ
Clouds	12,68	14,84	–	12,4	14,8	11,8	12,2	11,7
Concentric	1,36	1,2	–	1,17	1,04	1,7	2,8	1,5
Gaussian2 S=2	28,12	35,12	–	27,83	35,64	27,4	26,8	27,2
Gaussian7 S=2	14,35	15,68	–	14,36	15,76	15,9	15,3	11,5
Gaussian7 S=5	14,69	14,64	–	13,36	15,12	–	–	–
Iris S=2	6,53	6,67	5,33	5,59	6,67	4	4,3	6,7
Iris S=4	4,27	6,67	5,73	6,13	6,67	–	–	–
Phoneme S=2	14,12	11,51	13,7	15,79	14,47	12,3	16,3	16,4
Phoneme S=5	13,61	11,62	13,19	14,82	12,62	–	–	–
Satimage S=2	10,06	10,13	9,15	10,82	10,79	9,9	12,3	11,4
Satimage S=5	10,11	–	9,1	10,64	–	–	–	–
Texture S=2	0,82	0,76	1,13	0,82	0,80	1,9	2,0	3,1
Texture S=5	0,73	–	1,07	0,74	–	–	–	–

Классификация текстов и изображения. Традиционные подходы к классификации текстов используют в качестве элементов их векторных представлений функции частот встречаемости слов. Для сокращения размерности векторов могут применяться методы отбора информативных признаков [4], однако даже в случае отсутствия учета зависимостей между признаками вычислительные затраты на отбор квадратично растут с их числом. Нами предлагается для сокращения размерности применить распределенные представления. Для этого каждому слову поставим в соответствие N-мерный код с m случайно расположенными единицами. Распределенное N-мерное представление текста будем формировать суммой векторов слов с последующей пороговой операцией, либо контекстно-зависимым прореживанием *CDT* (см. в [3]), и далее использовать в задаче классифика-

ции. Тестирование проводилось на текстах коллекции *Reuters*-21578 [3] с помощью *SVM*. Для категорий *TOP*-10 результат *BEP* (точка перелома характеристики полнота-точность [3]) для исходного представления N=20000 составляет 0.913. Применение распределенных представлений с N^*=1000, m=2 позволило получить сравнимый результат 0.861, а использование *CDT* в ряде случаев повышало его на несколько процентов.

Аналогично формируемые распределенные представления исследовались при классификации изображений рукописных цифр базы *MNIST* [4]. Изображения кодировались путем выделения бинарных признаков. Наличие признака соответствовало комбинации белых и черных точек в некоторых позициях сетчатки. В результате кодирования формировалось "первичное" бинарное распределенное представление. Затем осуществлялось его преобразование во "вторичное" представление по той же методике, что и для текстовой информации, и классификация "вторичных" распределенных представлений.

Исследовалась классификация при сокращении размерности с N до N^* (таблица 2). Строка *sel* содержит число ошибок классификации, полученное с применением отбора признаков [4], а *distr* - с использованием распределенных представлений. Результаты для распределенных представлений значительно превышают результаты исходных представлений для тех же N (приведены в скобках в первой строке) и находятся на уровне результатов с отбором признаков.

Таблица 2

N (err)	5000 (667)	10000 (407)		50000 (195)			128000 (160)		
N^*	1000	1000	5000	1000	5000	10000	1000	5000	10000
sel	820	578	420	492	264	242	474	261	218
distr	904	727	415	632	274	213	826	264	204

Выводы. Исследованы разработанные бинарные распределенные представления векторных данных (числовых, текстовых, изображений) в задачах классификации. Проведен сравнительный анализ результатов для разных методов и задач с искусственными и естественными данными. Исследования показали, что полученные ранее аналитические выражения для характеристик кодов числовых векторов *RSC-Prager* позволяют выбирать их параметры, обеспечивающие высокие результаты в задачах классификации при использовании линейных классификаторов. Результаты, полученные предложенным персептроном с защитной полосой и одновременным обучением на не-

 ISSN 1562-9945

сколько классов, сопоставимы с результатами одного из лучших классификаторов *SVM* при значительном уменьшении времени обучения и распознавания. Результаты, полученные с ядрами *RSC-Prager*, также позволяют сократить время обучения для малых *S*.

Применение распределенного кодирования для представления бинарных признаков в задачах классификации текстов и изображений также позволило получить вычислительно эффективные решения при сохранении качества классификации. Перспективными направлением дальнейших исследований являются разработка вычислительно эффективных ядер *RSC* и *Prager*, а также распределенных представлений и ядер для более адекватного учета структурной информации во входных данных.

ЛИТЕРАТУРА

1. 1. *Duda R., Hart P., Stork D.* Pattern Classification, 2nd ed. - New York: John Wiley & Sons, 2000. - 680 p.
2. 2. *Слипченко С.В., Мисуно И.С., Рачковский Д.А.* Свойства кодирования числовых величин случайными гиперпрямоугольными рецептивными полями // Математические машины и системы. – 2005, № 4. – С. 15-29.
3. 3. *Мисуно И.С., Рачковский Д.А., Слипченко С.В., Соколов А.М.* Поиск текстовой информации с помощью векторных представлений // Проблемы программирования. – 2005, № 4. – С. 50–59.
4. 4. *Мисуно И.С., Рачковский Д.А., Слипченко С.В.* Экспериментальное исследование классификации рукописных цифр // Системные технологии. – 2005. – № 4 (39). – С. 110–133.
5. 5. *Vapnik V.N.* Statistical Learning Theory. – New York: John Wiley & Sons, 1998. – 768 p.
6. 6. *Мисуно И.С., Рачковский Д.А., Ревунова Е.Г., Слипченко С.В., Соколов А.М., Тетерюк А.Е.* Модульный программный нейрокомпьютер SNC: реализация и применение // УСиМ. – 2005, № 2. – С. 74–85.
7. 7. *Zhora D.* Evaluating Performance of Random Subspace Classifier on ELENA Classification Database // Artificial Neural Networks: Biological Inspirations – ICANN 2005. – Springer–Verlag Berlin Heidelberg. – 2005. – P. 343–349.

Получено 17.03.2006 г.