УДК 681.3

А.А. Шумейко, С.Л. Сотник, М.В. Лысак

ИСПОЛЬЗОВАНИЕ ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ В ЗАДАЧАХ КЛАССИФИКАЦИИ ТЕКСТОВ

Введение

Лавинообразное количество информации, вырабатываемое человечеством, привело к понятию автоматизации извлечения знаний –Data Mining. Это направление связано с широким спектром задач - от распознавания размытых образов до создания поисковых машин. Важной составляющей Data Mining является обработка текстовой информации. Такого рода задачи опираются на понятие классификации и кластеризации. Классификация заключается в определении принадлежности некоторого элемента (текста) одному из заранее созданных классов. Кластеризация подразумевает разбиение множества элементов (текстов) на кластеры, количество которых определяется локализацией элементов заданного множества в окрестностях некоторых естественных центров этих кластеров. Реализация задачи классификации изначально должна опираться на заданные постулаты, основные из которых – априорная информация о первичном множестве текстов и мера близости элементов и классов.

Постановка задачи классификации.

Будем использовать следующую модель задачи классификации.

Ω- множество объектов распознавания (пространство образов).

$\omega \in \Omega$ объект распознавания (образ).

$g(\omega):\Omega \to \Re$, $\Re=\{1,2,...,n\}$- индикаторная функция, разбивающая пространство образов Ω на n непересекающихся классов $\Omega^1,\Omega^2,...,\Omega^n$. Индикаторная функция неизвестна наблюдателю.

X — пространство наблюдений, воспринимаемых наблюдателем (пространство признаков).

$x(\omega):\Omega \to X$ — функция, ставящая в соответствие каждому объекту ω точку $x(\omega)$ в пространстве признаков. Вектор $x(\omega)$- это образ объекта, воспринимаемый наблюдателем.

© Шумейко А.А., Сотник С.Л., Лысак М.В., 2009

В пространстве признаков определены непересекающиеся множества точек Ξ[i]⊂X i=1,2…,n, соответствующих образам одного класса.

φ(x): X→ℜ решающее правило - оценка для g(ω) на основании x(ω), т.е. φ(x)= φ(x(ω)).

Пусть $x_\nu = x(\omega_\nu), \nu = 1,2,...,N$ доступная наблюдателю информация о функциях g(ω) и x(ω), но сами эти функции наблюдателю неизвестны. Тогда $(g_\nu, x_\nu), \nu = 1,2,...,N$ — есть множество прецедентов.

Задача заключается в построении такого решающего правила φ(x), чтобы распознавание проводилось с минимальным числом ошибок.

Основные направления исследования проблемы классификации.

Обычный случай — считать пространство признаков евклидовым, а качество решающего правила измеряют частотой появления правильных решений. Как правило, его оценивают, наделяя множество объектов Ω некоторой вероятностной мерой. Байесовский подход (см., например, [1]) исходит из статистической природы наблюдений. За основу берется предположение о существовании вероятностной меры на пространстве образов, которая либо известна, либо может быть оценена. Цель состоит в разработке такого классификатора, который будет правильно определять наиболее вероятный класс для пробного образа. Тогда задача состоит в определении “наиболее вероятного” класса. Байесовский подход основывается на предположении о существовании некоторого распределения вероятностей для каждого параметра. Недостатком этого метода является необходимость постулирования как существования априорного распределения для неизвестного параметра, так и знание его формы.

Использованию поиска соответствия предшествует построение множества статистик, в которых содержится количество текстов в данном классе и список используемых термов вместе со своими счетчиками.

Для определения подходящего класса текстов для заданного текста строится структура из неповторяющихся термов и их счетчиков - $(w_i, n(w_i))$.

 ISSN 1562-9945

Через M обозначим количество множества статистик. Классы, на принадлежность к которым проверяется текст, обозначим через $C_j (j = 0,...,M-1)$. Для каждого слова w_i из проверяемого текста, в каждой статистике находим это слово и соответствующий счетчик $n(w_i, C_j)$ (здесь j (j=0,1,..,M-1)-номер класса (элемент множества статистик)). Через $n(C_j)$ обозначим число текстов в j-м классе. Минимизация риска и вероятности ошибки эквивалентны разделению пространства признаков на n областей. Если области смежные, то они разделены поверхностью решения в многомерном пространстве. Для случая построения разделяющей поверхности предпочтительней использовать методы классификации отличные от Байесовской. Использование вероятностных характеристик определяется на распределение Гаусса, которое очень широко используется по причине вычислительного удобства и адекватности во многих случаях.

Если известно или с достаточным основанием можно считать, что плотность распределения функций правдоподобия $P(x|\Omega^i)$ является гауссовской, то применение классификатора Байеса приводит к тому, что образы, характеризующиеся нормальным распределением проявляют тенденцию к группированию вокруг среднего значения, а их рассеивание пропорционально среднеквадратическому отклонению σ. Вероятностные методы опираются на информацию о плотности распределения вероятностей каждого класса. К сожалению, в реальных задачах информация о плотности распределения отсутствует.

J.Rocchio [2] для решения задачи автоматической классификации объектов аэрокосмической съемки, предложил алгоритм TFIDF (term frequency / inverse document frequency), который состоит в следующем. Для каждого класса путем комбинации положительных и отрицательных решений принадлежности нормализованных векторов данному классу, строится центральный вектор этого класса c_i, а в качестве меры используется значение косинуса угла между проверяемым вектором и центральным вектором класса. Таким образом, задача нахождения подходящего класса для вектора d' сводится к решению задачи

$$\arg\max_{C_j} \cos\left(c_j, d'\right) = \arg\max_{C_j} \frac{\left\langle c_j, d' \right\rangle}{\left|c_j\right|\left|d'\right|}.$$

К сожалению, эффективность этого метода существенно зависит от экспертных оценок положительных и отрицательных решений, используемых при конструировании центрального вектора класса.

В 1974 г. вышла книга В.Н.Вапника и А.Я.Червоненкиса [3], положившая начало целой серии их работ в этой области. Предложенные авторами методы распознавания образов и статистическая теория обучения, лежащая в их основе, оказались, весьма успешными. Алгоритмы классификации и регрессии под общим названием SVM во многих случаях успешно заменили нейронные сети и в данное время применяются очень широко.

Идея метода основывается на предположении о том, что наилучшим способом разделения точек в n-мерном пространстве является n-1 плоскость (заданная функцией f(x)), равноудаленная от точек, принадлежащих разным классам. Метод опорных векторов (Support Vector Machine - SVM) относится к группе граничных методов. Эта группа методов определяет классы при помощи границ областей. Опорными векторами называются объекты множества, лежащие на границах областей. Классификация считается хорошей, если область между границами пуста. Однако сложность построения SVM-модели заключается в том, что чем выше размерность пространства, тем сложнее с ним работать, что существенно ограничивает использование SVM.

Основные результаты.

Для построения файла статистики последовательно обрабатываются все файлы словоформ $b^\nu, \nu = 0,..., M-1$, принадлежащие одному классу $B = \left\{b^\nu\right\}_{\nu=0}^{M-1}$. По множеству словоформ каждого обрабатываемого текста b^ν строится множество уникальных (неповторяющихся) словоформ и их счетчики - $\left(w_i^\nu, n_i^\nu\right)\left(i = 0,..., N^\nu - 1\right)$. Здесь N^ν - количество уникальных словоформ для текста b^ν. После этого данные для каждого файла отдельно нормируются

$$\overline{n}_i^\nu = \frac{n_i^\nu}{\sqrt{\sum_{j=0}^{N^\nu-1}\left(n_j^\nu\right)^2}}\left(i = 0,..., N^\nu - 1\right).$$

После этого, упорядочиваем все слова для каждого документа в одном и том же порядке (сам порядок слов не существенен, главное, чтобы слова в каждой из структур $\left(w_i^\nu, n_i^\nu\right)\left(i=0,...,N^\nu-1\right)$ шли в одном и том же порядке) и находим сумму всех векторов $n_i\left(B\right)=\sum_{j=0}^{M-1}\overline{n}_i^\nu\left(i=0,...,N\left(B\right)\right)$ (где N(B)- количество уникальных словоформ для класса B в целом) и нормируем ее единицей

$$\overline{n}_i\left(B\right)=\frac{n_i\left(B\right)}{\sqrt{\sum_{j=0}^{N(B)}\left(n_j\left(B\right)\right)^2}}.$$

Для полученной центральной точки класса формируем файл статистики, записывая в него значения $\left(w_i\left(B\right),\overline{n}_i\left(B\right)\right)\left(i=0,...,N\left(B\right)\right)$.

Для построения центрального вектора классов $\left\{B^\mu\right\}_{\mu=0}^{K-1}$, где каждый класс B^μ описывается своим центральным вектором $\left(w_i\left(B^\mu\right),\overline{n}_i\left(B^\mu\right)\right)\left(i=0,...,N\left(B^\mu\right)\right)$ нужно найти их сумму, просуммировав все координаты из всех суммируемых векторов для каждого значений словоформы, то есть для словоформы ω получаем координату

$$n\left(\omega\right)=\sum_{\mu=0}^{K-1}\left\{\overline{n}_i\left(B^\mu\right)\middle|\, w_i\left(B^\mu\right)=\omega, i=0,...N\left(B^\mu\right)\right\},$$

то есть, нужно составить список уникальных словоформ по всем центральным векторам классов $\left\{B^\mu\right\}_{\mu=0}^{K-1}$ и просуммировать их координаты. Результатом будет множество, состоящее их уникальных (неповторяющихся) словоформ и их координат

$$\left(w_i\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right), n_i\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)\right)\left(i=0,...,N\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)\right)$$

где $N\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)$ количество уникальных словоформ множества классов $\left\{B^\mu\right\}_{\mu=0}^{K-1}$. Остается пронормировать полученные координаты

$$\overline{n}_i\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)=\frac{n_i\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)}{\sqrt{\sum_{j=0}^{N\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)}\left(n_j\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)\right)^2}}$$

и полученный вектор $\left(w_i\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right),\overline{n_i}\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)\right)\left(i=0,...,N\left(\left\{B^\mu\right\}_{\mu=0}^{K-1}\right)\right)$ будет центральным вектором множества $\left\{B^\mu\right\}_{\mu=0}^{K-1}$.

Идеально сформированной классификацией векторного метода является такой набор классов $\left\{B^\mu\right\}_{\mu=0}^{K-1}$, для которого выполняется следующее условие $\forall b\in B^\mu,\mu=0,...,K-1$ имеет место соотношение

$$\left\langle\overline{n}(b),\overline{n}\left(B^\mu\right)\right\rangle<\left\langle\overline{n}(b),\overline{n}\left(B^\nu\right)\right\rangle,\nu\neq\mu. \qquad (1)$$

Рассмотрим вектор Λ (управление) размерностью $N\left(B^\mu\right)$, координаты которого принимают только одно из двух допустимых значений

$$\lambda_i=\begin{cases}0,\\1.\end{cases}$$

Через Λb обозначим прямое произведение векторов Λ и b, то есть

$$\Lambda b=\left(\lambda_0\overline{n}_0(b),\lambda_1\overline{n}_1(b),...,\lambda_{N(B^\mu)}\overline{n}_{N(B^\mu)}(b)\right).$$

Управление Λ будем называть допустимым на классе $B^\mu=\left\{b^k\right\}_{k=0}^{M-1}$, если выполняется условие

$$\left\langle\overline{\Lambda\overline{n}}\left(b^k\right),\overline{\Lambda\overline{n}}\left(B^\mu\right)\right\rangle<\left\langle\overline{\Lambda\overline{n}}\left(b^k\right),\overline{n}\left(B^\nu\right)\right\rangle,\nu\neq\mu,k=0,1,...,M-1 \qquad (2)$$

Допустимое управление, для которого имеет место это соотношение и при этом $\sum_{k=0}^{M-1}\left(\Lambda b^k\right)^2\to\max$ называется оптимальным.

Если для $\nu\neq\mu$ множество допустимых управлений вырождено, то класс $B^\mu=\left\{b^k\right\}_{k=0}^{M-1}$ определен некорректно, то есть он неразделим с классом B^ν.

Задача нахождения оптимального управления классическими методами достаточно сложна, поэтому для ее решения мы применим генетические алгоритмы.

Основные принципы работы ГА заключены в следующей схеме:

1. Генерируем начальную популяцию из n хромосом λ_i.

2. Вычисляем для каждой хромосомы ее пригодность, то есть выполнение условия (2).

 ISSN 1562-9945

3. Выбираем пару хромосом-родителей с помощью одного из способов отбора.

4. Генерируем потомство выбранных родителей, используя генетические операторв, прежде всего еолссовер и мутацию.

5. Повторяем шаги 3–4, пока не будет сгенерировано новое поколение популяции, содержащее n хромосом.

6. Повторяем шаги 2–5, пока не будет достигнут критерий окончания процесса.

Критерием окончания процесса может служить заданное количество поколений или схождение популяции.

Существует несколько подходов к выбору родительской пары. Селекция состоит в том, что родителями могут стать только те особи, значение приспособленности которых не меньше пороговой величины, в нашем случае, среднего значения приспособленности по популяции. Такой подход обеспечивает более быструю сходимость алгоритма.

Оператор рекомбинации применяют сразу же после оператора отбора родителей для получения новых особей-потомков. Смысл рекомбинации заключается в том, что созданные потомки должны наследовать генную информацию от обоих родителей. Рекомбинацию бинарных строк принято называть кроссинговером (кроссовером) или скрещиванием. В нашем случае используется одноточечный кроссинговер (Single-point crossover), который моделируется следующим образом -пусть имеются две родительские особи с хромосомами $X = \{x_i, i \in \{0,...,L\}\}$ и $Y = \{y_i, i \in \{0,...,L\}\}$. Случайным образом определяется точка внутри хромосомы (точка разрыва), в которой обе хромосомы делятся на две части и обмениваются ими. После процесса воспроизводства происходят мутации (mutation). Данный оператор необходим для «выбивания» популяции из локального экстремума и препятствует преждевременной сходимости. Это достигается за счет того, что изменяется случайно выбранный ген в хромосоме.

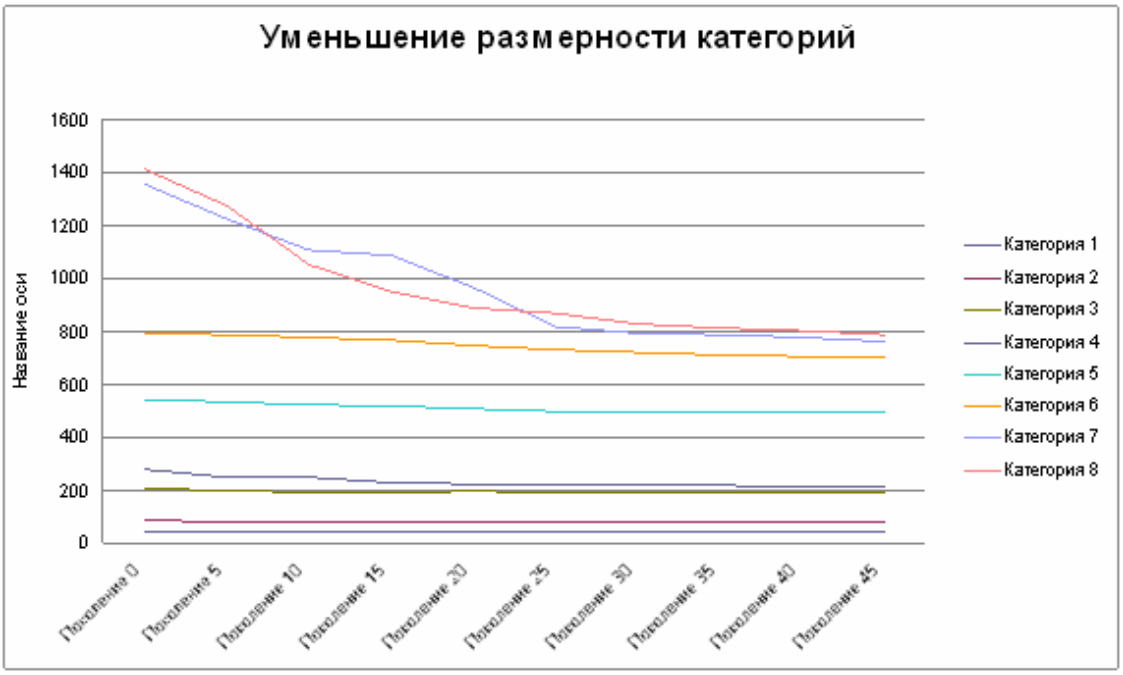

Рисунок 1 - Диаграмма уменьшения размерности категорий при использовании генетических алгоритмов

Для создания новой популяции можно использовать различные методы отбора особей. Нами использован элитарный отбор (Elite selection). Создается промежуточная популяция, которая включает в себя как родителей, так и их потомков. Члены этой популяции оцениваются, а за тем из них выбираются N самых лучших (пригодных), которые и войдут в следующее поколение.

Результат применения генетического алгоритма к задаче сокращения размерности класса, приведен на рисунке 1.

Заметим, что векторный метод в качестве критерия качества использует величину скалярного произведения ортов, таким образом, класс единичных векторов (документов) ограничен на сфере окружностью с центром в конце центрального вектора класса. Так как срезы сферы по окружности не могут плотно упаковать всю поверхность единичной сферы, то появляется множество точек (ортов), которые принципиально не могут попасть ни в один класс. Таким образом, возникает необходимость разбить множество точек на единичной сфере, так, чтобы элементы этого разбиения плотно упаковывали всю поверхность единичной сферы, то есть позволяли однозначно классифицировать любой документ.

Использование диаграмм Вороного в задаче классификации текстов.

 ISSN 1562-9945

Для любого центра системы $\{A\}$ можно указать область пространства, все точки которой ближе к данному центру, чем к любому другому центру системы. Такая область называется многогранником Вороного или областью Вороного. К многограннику Вороного обычно относят и его поверхность. В трехмерном пространстве область Вороного любого центра i системы $\{A\}$ есть выпуклый многогранник, в двумерном — выпуклый многоугольник. Формально многоугольники Вороного T_i определяются следующим образом:

$$T_i = \left\{ x \in R^2 : d\left(x, x_i\right) < d\left(x, x_j\right) \forall j \neq i \right\}$$

Построение аппроксимации опирается на фундаментальное свойство для произвольно заданных n точек множества S на плоскости. Для любого узла из n на плоскости существует множество натуральных соседей N. Понятие натуральных соседей тесно связано с разбиением области ячейками Вороного. Для непустой ячейки Вороного $V(R), R \subset S$ натуральные соседи для $r \in R$ - вершины треугольников Делоне, инцидентных к $V(R)$.

Двумерный многогранник Вороного показан на рисунке 2. Плоскости Вороного, которые породили грани у данного многогранника, называются образующими плоскостями Вороного, а соответствующие центры системы — геометрическими соседями данного центра i. Среди геометрических соседей различают основные (естественные) и не основные. Для первых середина отрезка, соединяющая его с центральным узлом, лежит на грани многогранника Вороного. Для вторых — вне грани и, следовательно, вне самого многогранника.

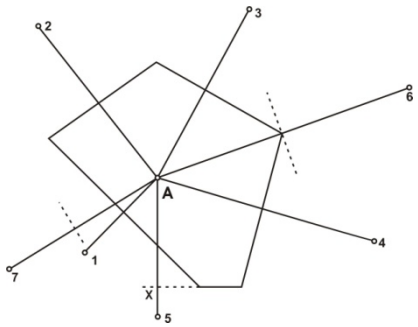

Рисунок 2 - Многогранник Вороного для центра i двумерной системы

Многогранники Вороного, построенные для каждого центра системы $\{A\}$, дают мозаику многогранников - разбиение Вороного (рисунок 3). Многогранники Вороного системы $\{A\}$ не входят друг в друга и заполняют пространство, будучи смежными по целым граням. Разбиение пространства на многогранники Вороного однозначно определяется системой $\{A\}$ и, наоборот, однозначно ее определяет.

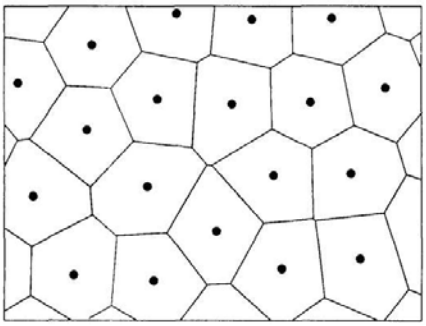

Рисунок 3 - диаграмма Вороного на плоскости

Используя конструкцию диаграмм Вороного применительно к точкам на многомерной единичной сфере, получаем разбиение всех ортов документов на естественные классы. Границами классов будут являться гиперплоскости, разделяющие сферические многогранники Вороного. При этом к одному классу будут относиться точки на единичной сфере (концы ортов документов), которые по отношению ко всем гиперплоскостям, ограничивающим данный класс, лежат с одной ее стороны, что и центральный вектор этого класса.

Проверка существующей классификации на корректность.

Пусть проверяются на корректность разбиения классы документов C_ν и C_μ. Для соответствующих орт (центральных векторов) $\widehat{C_\nu}$, $\widehat{C_\mu}$ строим вектор разности

$$\vec{\Delta}_{\nu,\mu} = \widehat{C}_\nu - \widehat{C}_\mu = \left\{\hat{n}^\nu\left(w_i\right) - \hat{n}^\mu\left(w_i\right)\right\}$$

и вектор суммы

$$\vec{\Xi}_{\nu,\mu} = \frac{1}{2}\left(\widehat{C}_\nu + \widehat{C}_\mu\right) = \frac{1}{2}\left\{\hat{n}^\nu\left(w_i\right) + \hat{n}^\mu\left(w_i\right)\right\}.$$

 ISSN 1562-9945

Конец вектора полусуммы численно совпадает с координатами этого вектора. Обозначим его через $\Xi_{\nu,\mu}$. Проведем через точку $\Sigma_{\nu,\mu}$ плоскость с нормальным вектором $\vec{\Delta}_{\nu,\mu}$

$$\Omega_{\nu,\mu} = \left\langle \vec{\Delta}_{\nu,\mu} \cdot \left(\mathrm{P} - \Xi_{\nu,\mu} \right) \right\rangle = 0. \qquad (3)$$

Эта плоскость разделяет классы. Для того, чтобы метод корректно разделял классы, нужно, чтобы все точки (документы) одного класса находились с одной стороны плоскости, то есть, если $b \in C_\nu$, то

$$\left\langle \vec{\Delta}_{\nu,\mu} \cdot \left(\hat{C}_\nu - \Xi_{\nu,\mu} \right) \right\rangle \left\langle \vec{\Delta}_{\nu,\mu} \cdot \left(\hat{b} - \Xi_{\nu,\mu} \right) \right\rangle \geq 0.$$

Точки, в которых это условие не выполняется нужно рассмотреть на вопрос принадлежности к категории C_μ.

Возможна ситуация, когда категории имеют непустое пересечение. Например, категория «АВТОР» содержит документы по математике (этого автора) и категория «НАУКА» содержит документы по математике того же автора. В этом случае нужно выделить непустое пересечение этих категорий и, в дальнейшем, либо эту категорию локализовать, либо проводить адресацию в обе категории.

Для решения этой проблемы предлагается следующий метод. Рассмотрим категории C_ν и C_μ. Разделим их плоскостью (3), и все точки, лежащие с одной стороны, соберем в новые две категории C_ν^* и C_μ^*.

Пусть

$$d\left(B, \Omega_{\nu,\mu}\right) = \frac{\left| \left\langle \vec{\Delta}_{\nu,\mu} \cdot \left(B - \Xi_{\nu,\mu} \right) \right\rangle \right|}{\left| \vec{\Delta}_{\nu,\mu} \right|}$$

расстояние от точки $B = \{b_i\}$ до плоскости $\Omega_{\nu,\mu}$.

Если выполняется условие (то есть, после отсечения данных обе категории отодвигаются друг от друга)

$$\begin{cases} d\left(C_\nu^*, \Omega_{\nu,\mu}\right) - d\left(C_\nu, \Omega_{\nu,\mu}\right) > 0, \\ d\left(C_\mu^*, \Omega_{\nu,\mu}\right) - d\left(C_\mu, \Omega_{\nu,\mu}\right) > 0, \end{cases}$$

то категории C_ν и C_μ имеют непустое пересечение $\tilde{C}$, которое можно определить следующим образом, $b \in \tilde{C}$ если $b \in C_\nu$ и при этом

$$\left\langle \vec{\Delta}_{\nu,\mu} \cdot \left(\hat{C}_{\nu} - \Xi_{\nu,\mu} \right) \right\rangle \left\langle \vec{\Delta}_{\nu,\mu} \cdot \left(\hat{b} - \Xi_{\nu,\mu} \right) \right\rangle < 0$$

или, если $b \in C_{\mu}$ и при этом

$$\left\langle \vec{\Delta}_{\nu,\mu} \cdot \left(\hat{C}_{\mu} - \Xi_{\nu,\mu} \right) \right\rangle \left\langle \vec{\Delta}_{\nu,\mu} \cdot \left(\hat{b} - \Xi_{\nu,\mu} \right) \right\rangle < 0 .$$

Естественно, что для метода, построенного на диаграммах Вороного также актуальна задача сокращения размерности классов. Для этой цели, так же, как и в предыдущем случае, использовались генетические алгоритмы.

Сравнительный анализ применения различных методов классификации к тестовой базе документов Reuters приведен на следующих диаграммах.

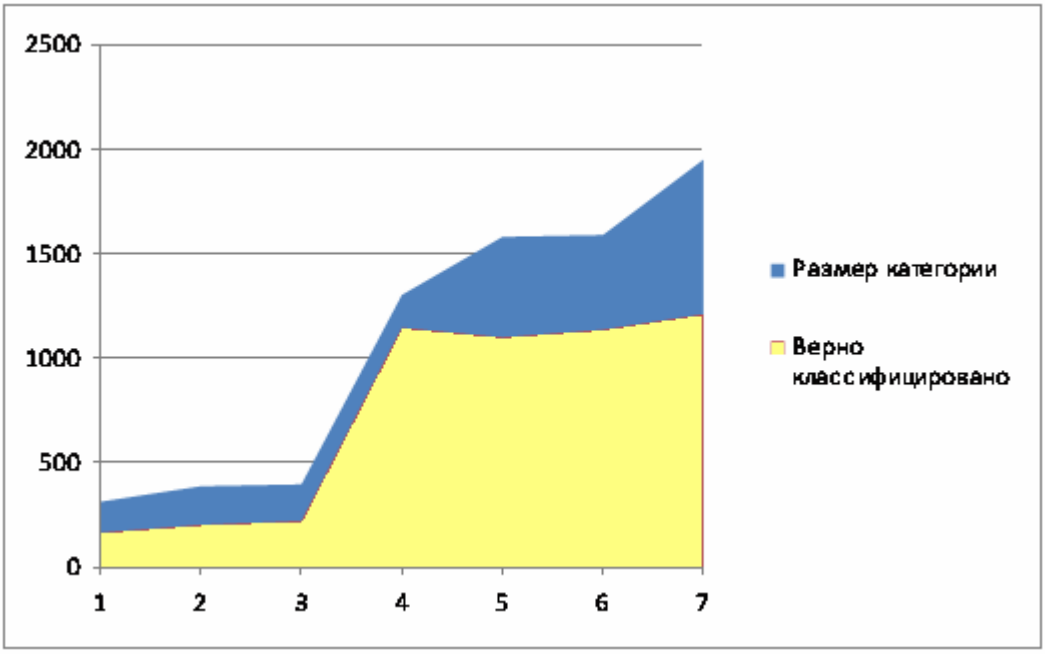

Рисунок 4 - Результат применения алгоритма Байеса

 ISSN 1562-9945

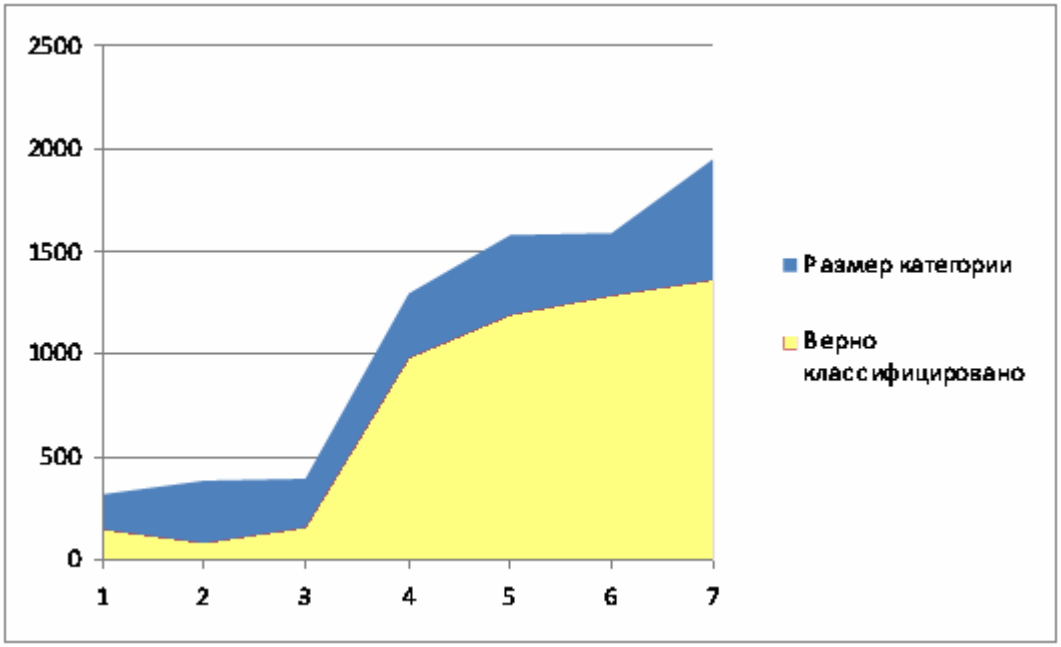

Рисунок 5 - Результат применения векторного алгоритма

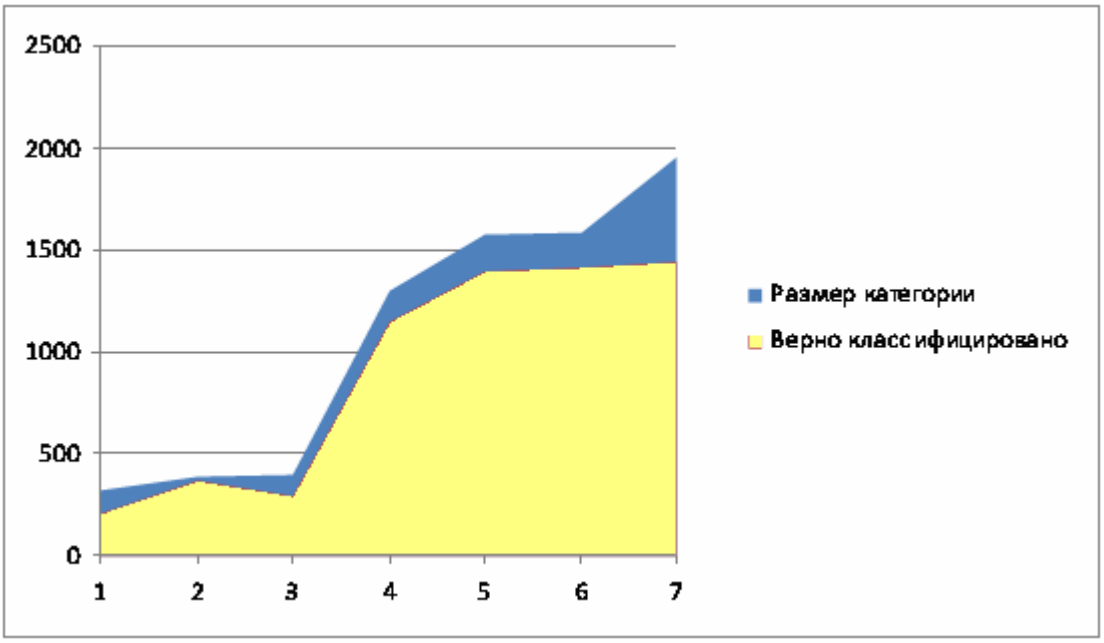

Рисунок 6 - Результат применения алгоритма основанного на диаграммах Вороного

Выводы.

Результаты работы показали, что для сокращения размерности векторов классов документов достаточно эффективно использовать генетические алгоритмы с хромосомами длиной равной количеству ненулевых координат центрального вектора и бинарными генами. Предложено минимизацию генотипа проводить из условия

минимальности воздействия на класс при обеспечении заданной степени локализации класса. Для тестовой базы Rambler при условии попадания в класс не менее 90% документов, удалось сократить размерность классов от 10% до 50%.

ЛИТЕРАТУРА

1. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и Data Mining.- СПб.: БХВ-Петербург, 2004.- 336 с.
2. Rocchio O. “Relevance Feedback in Information Retrieval”, in Salton: The SMART Retrieval System: Experiments in Automatic Document Processing, Chapter 14, P.313-323, Prentice-Hall, 1971.
3. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. — М.: Наука, 1974. — 416 с.
4. Шумейко А.А., Сотник С.Л., Лысак М.В. Использование генетических алгоритмов в задачах классификации текстов// Тези доповідей на VI міжнародній науково-практичній конференції «Математичне та програмне забезпечення інтелектуальних систем» (MPZIS-2008), Дніпропетровськ, 2008.- С. 345-346.

Одержано 05.12.2008р.

 ISSN 1562-9945