Институт информационных и вычислительных технологий
МОН РК

Казахский Национальный Университет имени аль-Фараби

Университет Туран

МАТЕРИАЛЫ

II Международной научной конференции

«Информатика и прикладная математика»

«Информатика және қолданбалы математика»

«Computer Science and Applied Mathematics»

(ЧАСТЬ II)

27-30 сентября 2017 года, Алматы, Казахстан

Алматы 2017

ПОЛУКОНТРОЛИРУЕМОЕ ОБУЧЕНИЕ НА ОСНОВЕ КЛАСТЕРНОГО АНСАМБЛЯ

Бериков В.Б.[1,2] , Амиргалиев Е.Н. [3,4], Черикбаева Л.Ш [4].
e-mail: lyailya_sh@mail.ru

[(1)] Институт математики им. С.Л. Соболева СО РАН, г. Новосибирск
[(2)] Новосибирский государственный университет, г. Новосибирск
[(3)] Институт информационных и вычислительных технологий КН МОН РК, г. Алматы
[(4)] Казахский национальный университет имени аль-Фараби, г. Алматы

Аннотация

В работе рассмотрен один из вариантов постановки задачи распознавания образов - задача полуконтролируемого обучения. Были разработаны алгоритмы CASVM и CANN для решения этой задачи. Они основываются на сочетании методов коллективного кластерного анализа и ядерных методов классификации. Предложена вероятностная модель классификации с использованием кластерного ансамбля. В рамках модели исследовано поведение вероятности ошибки алгоритма CANN. Сформулированы предположения, при выполнении которых вероятность ошибки стремится к нулю. Проведено экспериментальное исследование предложенного алгоритма на гиперспектральном изображении. Показано, что алгоритм CASVM более устойчив к шуму, чем стандартный метод опорных векторов SVM.

Ключевые слова*: распознавания, классификация, гиперспектральное изображение, алгоритм, полуконтролируемого обучения*

1. Введение

Задача распознавания образов состоит в классификации объектов по нескольким классам (образам). Каждый объект характеризуется конечным набором признаков. Классификация происходит на основании прецедентов - объектов, для которых мы знаем классы, которым они принадлежат. В основной постановке задачи классы известны для всех объектов выборки, после чего подаются новые объекты, которые требуется наиболее точно отнести к какому-то классу (классификация с учителем, Supervised learning). В данной работе рассматривается один из вариантов постановки задачи распознавания образов - задача полуконтролируемого обучения (Semi-supervised learning). В этой задаче для некоторых объектов исходной выборки известны классы, для некоторых неизвестны. Эта задача актуальна по следующим причинам:

- неразмеченные данные дешевы;
- размеченные данные часто бывает сложно получить;
- использование неразмеченных данных совместно с небольшим

количеством размеченных может обеспечить значительный прирост качества обучения.

Существует множество алгоритмов и подходов к решению задачи полуконтролируемого обучения [1]. Цель данной работы заключается в разработке нового подхода для решения задачи полуконтролируемого обучения, его теоретическом и экспериментальном обосновании. Новизна работы состоит в сочетании алгоритмов коллективного кластерного анализа [2,3] и ядерных методов классификации (на примере метода опорных векторов, *SVM* [4] и метода ближайшего соседа, *NN*), а также в теоретическом анализе ошибки предложенного метода. Далее будет описана математическая постановка задачи, сделан обзор некоторых методов кластерного анализа и ядерных методов классификации, будет изложена идея предлагаемого метода и проведено его теоретическое и экспериментальное обоснование.

2. Математическая постановка задачи полуконтролируемого обучения

Пусть имеется генеральная совокупность объектов распознавания X и конечное множество меток классов Y. Все объекты описываются признаками. Под признаком объекта понимается отображение $f: X \to D_f$, где D_f — множество значений признака.

В зависимости от D_f признаки делятся на типы:

- Бинарный признак: $D_f = \{0,1\}$
- Количественный признак: $D_f = R$
- Номинальный признак: D_f — конечное множество
- Порядковый признак: D_f — конечное упорядоченное множество

При заданных признаках $f_1,..,f_m$ вектор $x = (f_1(\alpha),..,f_m(\alpha))$ называется признаковым описанием объекта $\alpha \in X$. Далее мы отождествляем объект и его признаковое описание. В задаче полуконтролируемого обучения на вход подается выборка $X_N = \{x_1,..,x_N\}$ объектов из X. В этой выборке присутствуют объекты двух типов:

- $X_c = \{x_1,..,x_k\}$ - размеченные объекты с заданными классами, которым они принадлежат: $Y_c = \{y_1,..,y_k\}$
- $X_u = \{x_{k+1},..,x_N\}$ - неразмеченные объекты

В различных вариантах постановки задачи требуется либо провести т.н. индуктивное обучение — построить алгоритм классификации $a: X \to Y$, который будет, минимизируя вероятность ошибки, сопоставлять классы объектам их X_u, а также новым объектам X_{test}, которые были недоступны на момент построения алгоритма, либо требуется провести трансдуктивное

обучение - получить метки классов только для объектов из X_u с минимальной ошибкой. В данной работе рассматривается второй вариант постановки задачи.

Далее будет приведен пример, который показывает, чем полуконтролируемое обучение отличается от классификации с учителем.

Пример: На вход подаются размеченные объекты $X_c = \{x_1, .., x_k\}$ с классами $Y_c = \{y_1, .., y_k\}$, где $y_i \in \{0,1\}, i = 1, .., k$. Объекты имеют два измеренных признака, их расположение в пространстве показано на Рисунке 1.

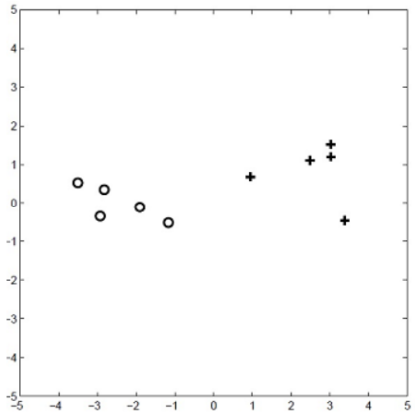


Рисунок 1. Размеченные объекты X_c. Разные метки соответствуют разным классам

Также на вход подаются неразмеченные данные $X_u = \{x_{k+1}, .., x_N\}$. Их расположение показано на Рисунке 2.

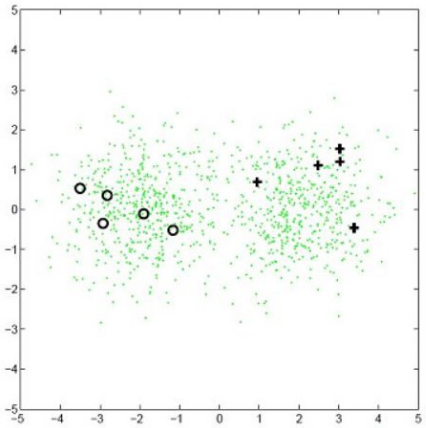


Рисунок. 2. Размеченные объекты X_c вместе с неразмеченными объектами X_u

Предположим, что дана выборка из смеси нормальных распределений и оценим плотности классов по всему набору данных и только по размеченным данным, после чего построим разделяющие кривые. Тогда из Рисунка 3 видно, что качество классификации при использовании полного набора данных выше.

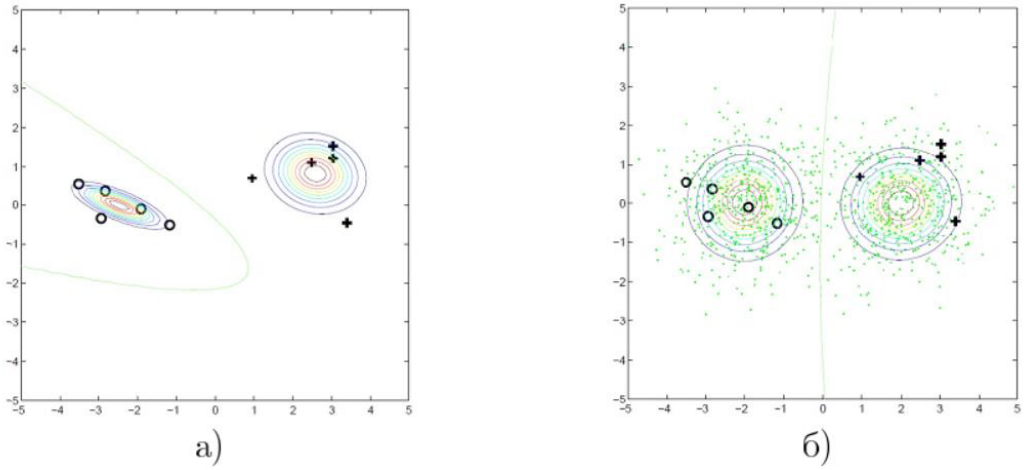


Рисунок 3. Восстановленные плотности классов: а) - по размеченным данным; б) - по неразмеченным данным

3. Коллективные решения в кластерном анализе

3.1. О причинах развития коллективного подхода

Задачей кластерного анализа является разбиение выборки на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер представлял группу похожих объектов, а объекты в разных кластерах существенно различались. Решение задачи кластеризации не является однозначным по нескольким причинам:

- Не существует наилучшего критерия качества кластеризации. Известно большое число разумных эвристических критериев и алгоритмов, не имеющих явно заданного критерия, но осуществляющих достаточно качественную кластеризацию;
- Число кластеров очень часто неизвестно заранее и устанавливается либо вручную, либо в ходе работы алгоритма;
- Результаты кластеризации очень сильно зависят от метрики, которая выбирается экспертом и специфики прикладной области.

Кроме того, алгоритмы кластерного анализа не универсальны: каждый алгоритм имеет свою специфическую область применения. Например, некоторые алгоритмы направлены на «шарообразные» структуры данных, другие на «ленточные» кластеры и т.д. На основании этих особенностей был предложен коллективный подход к кластерному анализу. В настоящее время

коллективный подход показывает наилучшие результаты по сравнению с отдельными алгоритмами [5] и позволяет использовать преимущества и особенности сразу нескольких алгоритмов. Существует несколько вариантов получения коллективного решения [6] задачи кластерного анализа: использование т.н. матрицы усредненных попарных различий, максимизация степени согласованности решений (с помощью исправленного индекса Ранда, нормализованной взаимной информации и т.д.), применение графовых методов. В предлагаемом в данной работе алгоритме используется матрица усредненных попарных различий.

3.1.2. Матрица усредненных попарных различий

Для построения матрицы усредненных попарных различий проводится кластеризация всех имеющихся объектов $X = \{x_1, .., x_N\}$ коллективом различных алгоритмов $\mu_1, .., \mu_M$ кластерного анализа. Каждый алгоритм дает L_m вариантов разбиения, $m = 1, .., M$. По результатам работы алгоритмов составляется матрица H усредненных попарных различий объектов из X . Элементы матрицы равны:

$$h(i, j) = \sum_{m=1}^{M} \alpha_m \frac{1}{L_m} \sum_{l=1}^{L_m} h_{lm}(i, j), \qquad (1)$$

где $i, j \in \{1, .., N\}$ - номера объектов $(i \neq j), \alpha_m \geq 0$ - заданные веса такие, что $\sum_{m=1}^{M} \alpha_m = 1; h_{lm}(i, j) = 0$, если пара (i, j) принадлежит разным кластерам в l-ом варианте разбиения, полученного алгоритмом μ_m и 1, если принадлежит одному кластеру.

Веса α_m могут быть одинаковыми или, например, могут подбираться пропорционально индексу качества кластеризации. Оптимальный выбор весов исследуется в работе [2].

4. Ядерные методы классификации

Для решения задачи классификации широко распространены ядерные методы, в основе которых лежит т.н. «kernel trick». Для демонстрации сути этого «трюка» рассмотрим метод опорных векторов (*SVM*) - наиболее популярный ядерный метод классификации. SVM является бинарным классификатором, хотя существуют способы его доработки для осуществления мультиклассификации.

4.1. Бинарная классификация с помощью SVM

В задаче разделения на два класса (задаче бинарной классификации) на вход подается обучающая выборка объектов $X = \{x_1, .., x_n\}$ с классами $Y = \{y_1, .., y_n\}$, $y_i \in \{+1, -1\}$ при $i = 1, .., n$, где объекты представляют из себя точки

в m-мерном пространстве признаковых описаний. Требуется разделить точки гиперплоскостью размерности ($m-1$). В случае линейной разделимости классов существует бесконечное число разделяющих гиперплоскостей. Разумно будет выбрать гиперплоскость, расстояние от которой до обоих классов максимально. Оптимальной разделяющей гиперплоскостью называется гиперплоскость, которая максимизирует ширину разделяющей полосы между классами. Задача метода опорных векторов заключается в построении оптимальной разделяющей гиперплоскости. При этом лежащие на краю разделяющей полосы точки называются опорными векторами.

Известно, что гиперплоскость представима в виде $\langle w,x\rangle+b=0$, где $\langle\ ,\ \rangle$ — скалярное произведение, w — вектор, перпендикулярный к разделяющей гиперплоскости, а b — вспомогательный параметр. Метод опорных векторов строит решающую функцию в виде.

$$F(x)=sign(\sum_{i=1}^{n}\lambda_i c_i\langle x_i,x\rangle+b)$$

Важно отметить, что суммирование идет только по опорным векторам, для которых $\lambda_i\neq 0$. Объекты $x\in X$ с $F(x)=1$ будут отнесены к одному классу, а объекты с $F(x)=0$ к другому.

При линейной неразделимости классов можно осуществить преобразование $\varphi: X\to G$ исходного пространства объектов X к новому пространству G большей размерности. Новое пространство называется спрямляющим, в нем объекты уже могут быть линейно разделимы

Решающая функция $F(x)$ зависит от скалярных произведений объектов, а не от самих объектов непосредственно. В силу этого скалярные произведения $\langle x,x'\rangle$ можно заменить на произведения вида $\langle\varphi(x),\varphi(x')\rangle$ в пространстве G. в этом случае решающая функция $F(x)$ будет иметь вид:

$$F(x)=sign(\sum_{i=1}^{n}\lambda_i c_i\langle\varphi(x_i),\varphi(x)\rangle+b)$$

Функция $K(x,x')=\langle\varphi(x),\varphi(x')\rangle$ называется ядром. Переход от скалярных произведений к произвольным ядрам и есть «трюк с ядром».

Подбор ядра определяет переход в спрямляющее пространство и позволяет применять линейные алгоритмы классификации (в частности, *SVM*) к линейно неразделимой выборке.

4.2. Теорема Мерсера

В ядерных методах классификации широко известна теорема [7], которая устанавливает необходимое и достаточное условие на то, чтобы функция была ядром:

Теорема (Мерсер) *Функция* $K(x,x')$ *является ядром тогда и только тогда, когда она симметрична* $K(x,x')=K(x',x)$, *и неотрицательно определена: для любой конечной выборки* $X^p=(x_1,...x_p)$ *из* X *матрица* $K=(K(x_i,x_j))$ *размера* $p\times p$ *неотрицательно определена: для любого* $z\in R^p$

5. Предлагаемый метод

Идея метода состоит в построении матрицы похожести (1) всех объектов из подаваемой на вход выборки X. Эта матрица будет составляться путем применения разных алгоритмов кластеризации к X. Чем чаще пара объектов попадает в один и тот же кластер, тем более похожими друг на друга мы их будем считать. Будет предложено два возможных варианта предсказания классов неразмеченных объектов X_u с использованием матрицы похожести. Далее идея алгоритма будет описана более подробно. Справедлива следующая:

Теорема 1. *Пусть* $\mu_1,..,\mu_M$ *— алгоритмы кластерного анализа, каждый алгоритм дает* L_m *вариантов разбиения,* $m=1,..,M$, $h_{lm}(x,x')=0$, *если пара* (x,x') *объектов принадлежит разным кластерам в* l*-ом варианте разбиения, полученного алгоритмом* μ_m *и* 1, *если принадлежит одному кластеру.* $\alpha_m\geq 0$ *- заданные веса такие, что* $\sum_{m=1}^{M}\alpha_m=1$. *Тогда функция* $H(x,x')=\sum_{m=1}^{M}\alpha_m\frac{1}{L_m}\sum_{l=1}^{L_m}h_{lm}(x,x')$ *удовлетворяет условиям теоремы Мерсера*

Доказательство. Очевидно, функция $H(x,x')$ симметрична. Пусть C_r^{lm} - множество индексов объектов, принадлежащих r-ому кластеру, полученному m-ым алгоритмом в l-ом варианте разбиения. Покажем, что $H(x,x')$ неотрицательно определена.

Возьмем произвольный $z\in R^p$ и докажем, что $z^THz\geq 0$

$$z^THz=\sum_{i,j=1}^{p}\sum_{m=1}^{M}\alpha_m\frac{1}{L_m}\sum_{l=1}^{L_m}h_{lm}(i,j)z_iz_j=\sum_{m=1}^{M}\alpha_m\frac{1}{L_m}\sum_{l=1}^{L_m}\sum_{i,j=1}^{p}h_{lm}(i,j)z_iz_j=$$

$$\sum s_{m=1}^{M}\alpha_m\frac{1}{L_m}\sum_{l=1}^{L_m}(\sum_{i,j\in C_1^{lm}}z_iz_j+...+\sum_{i,j\in C_{K_{lm}}^{lm}}z_iz_j)=\sum_{m=1}^{M}\alpha_m\frac{1}{L_m}\sum_{l=1}^{L_m}((\sum_{iinC_1^{lm}}z_i)^2+...+(\sum_{i\in C_{K_{lm}}^{lm}}z_i)^2)\geq 0.$$

Таким образом, функция $H(x,x')$ может быть использована в качестве ядра в ядерных методах классификации, в частности, в методе опорных векторов (SVM) и в методе ближайшего соседа (NN). Далее предлагается два варианта алгоритма, реализующего предлагаемый подход:

Алгоритм CASVM

Вход: объекты X_c с заданными классами Y_c и объекты X_u, число алгоримтов кластеризации M, число кластеризаций L_m каждым алгоритмом $\mu_m, m = 1,...,M$.

Выход: классы объектов X_u.

1. Провести кластеризацию объектов $X_c \cup X_u$ алгоритмами $\mu_1,..,\mu_M$ кластерного анализа, получив L_m вариантов разбиения от каждого алгоритма $\mu_m, m = 1,...,M$.

2. Вычислить матрицу H на $X_c \cup X_u$ по формуле (?).

3. Обучить *SVM* на размеченных данных X_c, используя матрицу H в качестве ядра.

4. С помощью *SVM* предсказать классы для неразмеченных объектов X_u.

Конец алгоритма

Алгоритм CANN

Вход: объекты X_c с заданными классами Y_c и объекты X_u, число алгоримтов кластеризации M, число кластеризаций L_m каждым алгоритмом $\mu_m, m = 1,...,M$.

Выход: классы объектов X_u.

1. Провести кластеризацию объектов $X_c \cup X_u$ алгоритмами $\mu_1,..,\mu_M$ кластерного анализа, получив L_m вариантов разбиения от каждого алгоритма $\mu_m, m = 1,...,M$.

2. Вычислить матрицу H на $X_c \cup X_u$ по формуле (1).

3. Использовать метод *NN*: каждому неразмеченному объекту $x \in X_u = \{x_{k+1},..,x_N\}$ сопоставить класс наиболее похожего в смысле $H(x,x')$ размеченного объекта $x' \in X_c = \{x_1,..,x_k\}$. Формальная запись: $x_i = \arg\max_{j=1,..,k} H(x_i,x_j),\ i = k+1,..,N$.

Конец алгоритма

Отметим, что в предложенных алгоритмах не требуется хранить в памяти матрицу H размера $N \times N$ целиком: достаточно хранить матрицу

кластеризаций размера $N \times L$, где $L = \sum_{l=1}^{M} L_m$, в этом случае матрицу H можно вычислять динамически. В прикладных задачах, как правило, $L << N$, например, при работе с пикселями изображений.

6. Теоретическое исследование качества алгоритма CANN

Напомним постановку задачи: на вход подается выборка $X_N = \{x_1, .., x_N\}$ объектов. В этой выборке присутствуют объекты двух типов:

- $X_c = \{x_1, .., x_k\}$ - размеченные объекты с заданными классами $Y_c = \{y_1, .., y_k\}$, $I_c = \{1, ..., k\}$ - индексы этих объектов
- $X_u = \{x_{k+1}, .., x_N\}$ — неразмеченные объекты, $I_u = \{k+1, ..., N\}$ - индексы этих объектов

Для упрощения обоснования предположим, что классы объектов не пересекаются, и проведем L кластеризаций одним алгоритмом μ кластерного анализа со случайными параметрами $\Omega_1, ..., \Omega_L$.

Введем следующие обозначения для $i, j \in I_u$:

$$h_l(x_i, x_j) = \begin{cases} 1, \text{алгоритм } \mu \text{ в варианте } l \text{ объединил пару } x_i, x_j \text{ в один кластер} \\ 0, \text{иначе;} \end{cases}$$

а также величины $L_1(i, j) = \sum_{l=1}^{L} h_l(x_i, x_j)$, $L_0(i, j) = L - L_1(i, j)$, которые представляют собой число вариантов кластеризаций, в которых алгоритм проголосовал за объединение пары x_i, x_j, или против объединения, соответственно.

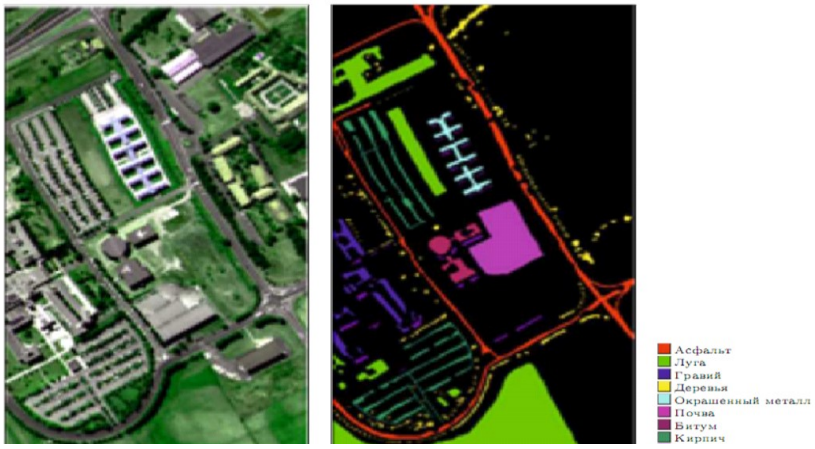
Пусть $Y(x)$ - скрытые от нас истинные метки классов неразмеченных объектов $x \in X_u$. Введем для $i, j \in I_u$ величину:

$$z(x_i, x_j) = \begin{cases} 1, \text{если } Y(x_i) = Y(x_j) \\ 0, \text{если } Y(x_i) \neq Y(x_j). \end{cases}$$

В методе ближайшего соседа (NN) для всех $i \in I_u$ мы присваиваем метке y_i значение y', где y' - это класс объекта $x' = \underset{x_j \in X_c}{arg\ max}\, H(x_i, x_j)$

Справедлива

Теорема 2. Пусть $\forall l \in \{1, ..., L\}, P[h_l(x_i, x_j) = 1 \mid z(x_i, x_j) = 1] > \frac{1}{2}$, $P[h_l(x_i, x_j) = 0 \mid z(x_i, x_j) = 0] > \frac{1}{2}$ и $L_0(i, j) = const \quad \forall i, j \in I_u$. Тогда в алгоритме CANN для объекта $x_i \in X_u$ вероятность неверной классификации $P_{er}(x_i) = P[Y(x_i) \neq y'] \to 0$ при $L \to \infty$.



а) *б*)

Рис. 4. Гиперспектральное изображение Pavia University scene (RGB композит) (а) и размеченные данные (б).

Доказательство данного свойства несколько громоздко и будет приведено в полной версии статьи.

Теорема показывает, что вероятность ошибки классификации алгоритмом *CANN* стремится к нулю при предположениях о том, что классы объектов не пересекаются и что используемые алгоритмы кластерного анализа правильно относят пары объектов к одному или разным кластерам с вероятностью более $1/2$, т.е., что они действуют не наугад.

7. Экспериментальное исследование

Обычное RGB -изображение содержит три канала: значения насыщенности по каждому из трех цветов. В некоторых случаях этого недостаточно, чтобы получить полную информацию о характеристиках снимаемого объекта. Для получения данных о неразличимых человеческим глазом свойствах объектов используется гиперспектральная съемка.

Для экспериментального исследования разработанного алгоритма было использовано изображение Pavia University scene [8] размером 610 на 340 пикселей, которое содержит 103 спектральных канала. Пространственное разрешение снимка составляет 1.3 м. На рисунке 4а) показан RGB-композит изображения (каналы 40, 50 и 70), а на рисунке 4б) приведено эталонное разбиение изображения на тематические классы.

Отметим, что на снимке имеются неразмеченные пиксели, которые не отнесены ни к одному из девяти классов. Данные пиксели были исключены из рассмотрения при анализе.

При экспериментальном исследовании алгоритма 1% пикселей, отобранных случайным образом для каждого класса, составили размеченную выборку; оставшиеся были включены в неразмеченную часть. Для изучения

влияния шума на качество работы алгоритма, случайно отобранные $r\%$ значений спектральных яркостей пикселей в разных каналах подвергались искажающему воздействию: соответствующее значение x заменялось величиной, выбранной случайным образом из интервала $[x(1-p),\ x(1+p)]$, где r, p - заданные параметры. Зашумленная таблица данных, содержащая значения спектральных яркостей пикселей по всем каналам, подавалась на вход алгоритма CASVM, а котором в качестве базового алгоритма для построения кластерного ансамбля был выбран алгоритм K-средних. Различные варианты разбиения получались варьированием числа кластеров в интервале $[30, 30+L]$, где L было равно 120. Кроме того, для построения каждого варианта решения случайным образом выбирались каналы, число которых было задано двум. Для ускорения работы алгоритма K-средних и получения более разнообразных вариантов группировки, число его итераций было ограничено значением 1.

Поскольку предложенный алгоритм реализует идею обучения метрике расстояния (distance metric learning), было бы естественно провести его сравнение с аналогичным алгоритмом (нашем случае - методом опорных векторов SVM), использующим стандартную евклидову метрику, в аналогичных условиях (выбирались параметры алгоритма, рекомендуемые по умолчанию в среде Матлаб). В таблице 1 показаны значения точности классификации неразмеченных пикселей изображения Pavia University scene для некоторых значений параметров зашумленности. Время работы алгоритма составило около 2 мин на двухъядерном процессоре Intel Core i5 с тактовой частотой 2.8 ГГц и объемом оперативной памяти 4 Гбайт. Как видно из таблицы, алгоритм CASVM обладает большей устойчивостью к шуму, чем алгоритм SVM.

Таблица 1. Точность алгоритмов CASVM и SVM при различных значениях параметров шума

Параметры шума r, p	**0%, 0**	**10%, 0.1**	**20%, 0.2**	**30%, 0.3**
CASVM	0.82	0.80	0.78	0.77
SVM	0.83	0.75	0.66	0.64

Заключение

В работе рассмотрен один из вариантов постановки задачи распознавания образов - задача полуконтролируемого обучения. Были разработаны алгоритмы *CASVM* и *CANN* для решения этой задачи. Они основываются на сочетании методов коллективного кластерного анализа и ядерных методов классификации.

Предложена вероятностная модель классификации с использованием кластерного ансамбля. В рамках модели исследовано поведение вероятности

ошибки алгоритма *CANN*. Сформулированы предположения, при выполнении которых вероятность ошибки стремится к нулю.

Проведено экспериментальное исследование предложенного алгоритма на гиперспектральном изображении. Показано, что алгоритм CASVM более устойчив к шуму, чем стандартный метод опорных векторов SVM.

Список литературы

[1] *Zhu X. Semi-supervised learning literature survey / Tech. Rep.* (Department of Computer Science, Univ. of Wisconsin, Madison, 2008), no. 1530.

[2] *Berikov V.B.* Weighted ensemble of algorithms for complex data clustering // Pattern Recognition Letters. 2014. Vol. 38. P. 99-106.

[3] *Berikov V., Pestunov I.* Ensemble clustering based on weighted co-association matrices: Error bound and convergence properties // Pattern Recognition. 2017. Vol. 63. P. 427-436.

[4] *Вапник В. Н.* Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979. 448 с.

[5] *Topchy A., Law M., Jain A., Fred A.* Analysis of consensus partition in cluster ensemble // Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04). 2004. P. 225-232.

[6] *Vega-Pons S., Correa-Morris J., Ruiz-Shulcloper J.* Weighted cluster ensemble using a kernel consensus function // LNAI. 2008. Vol. 5197. P. 195-202.

[7] *Mercer J.* Functions of positive and negative type and their connection with the theory of integral equations / Philos. Trans. Roy. Soc. London .—1909.

[8] Hyperspectral Remote Sensing Scenes. http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing\Scenes (дата обращения: 20.07.2017).

ПОФОНЕМНОЕ РАСПОЗНАВАНИЕ КАЗАХСКОЙ РЕЧИ С ИСПОЛЬЗОВАНИЕМ РАЗЛИЧНЫХ НАБОРОВ ФОНЕМ И ФОНЕТИЧЕСКИХ ТРАНСКРИПЦИЙ

Карабалаева М.Х., Есенбаев Ж.А.

muslima.karabalayeva@nu.edu.kz, zhyessenbayev@nu.edu.kz

National Laboratory Astana

Аннотация. *В данной работе были проведены эксперименты по пофонемному распознаванию слитной казахской речи с использованием различных наборов фонем и фонетических транскрипций. Целью экспериментов является исследование влияния фонетической системы языка,*

Содержание