**В. Б. БЕРИКОВ^{1,2}, Е. Н. АМИРГАЛИЕВ³, Л. Ш. ЧЕРИКБАЕВА⁴,
Б. А.* ТУЛЕГЕНОВА⁵**

¹Институт математики им. С.Л. Соболева СО РАН, Россия

²Новосибирский государственный университет, Россия

³Институт информационных и вычислительных технологий
КН МОН РК, Казахстан

⁴Казахский национальный университет имени аль-Фараби, Казахстан

⁵Казахский национальный исследовательский технический университет
им. К.И. Сатпаева, Казахстан

E-mail: berikov@math.nsc.ru; amir_ed@mail.ru; lyalya_sh@mail.ru;
tulegenova_bakhit@mail.ru

КЛАССИФИКАЦИЯ ПРИ НЕПОЛНОЙ ОБУЧАЮЩЕЙ ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ КЛАСТЕРНОГО АНСАМБЛЯ И МАЛОРАНГОВОЙ МАТРИЧНОЙ ДЕКОМПОЗИЦИИ

В работе решается задача распознавания образов в случае неполной обучающей информации. Предлагается метод решения, в котором сочетается регуляризация лапласиана графа сходства и коллективный кластерный анализ. Используется малоранговое представление матрицы коассоциации кластерного ансамбля, что позволяет ускорить вычисления и сэкономить память. Численные эксперименты на тестовых задачах и реальном гиперспектральном изображении демонстрируют эффективность предложенного метода, в том числе при наличии зашумленных данных.

Ключевые слова: распознавания, классификация, ансамбль, коассоциативная матрица, коллективный кластерный анализ.

В задаче распознавания образов требуется провести классификацию объектов выборки, которые описываются некоторым набором признаков. При этом необходимо получить оптимальное значение определенного критерия качества распознавания (например, минимизировать оценку вероятности ошибки). Классификатор формируется на основании обучающей выборки, состоящей из прецедентов – объектов, для которых известны образы, к которым они принадлежат.

В данной работе рассматривается один из вариантов постановки задачи распознавания образов – задача частично контролируемой классификации (*semi-supervised classification*). В этой задаче лишь для части объектов исходной выборки известны метки классов; требуется классифицировать либо имеющиеся неразмеченные объекты, либо сформировать решающее правило для распознавания новых объектов.

Существует достаточно большое число подходов и алгоритмов решения задачи частично контролируемой классификации [1]. Одно из первых решений этой задачи было предложено Н.Г. Загоруйко [2]. В настоящее время широко используются эвристические алгоритмы самообучения, вероятностные методы, трансдуктивный метод опорных векторов, а также теоретико-графовые алгоритмы (Graph Laplacian Regularization) [3-6].

Ранее в работах [7,8] был предложен, исследован новый подход к решению данной задачи, основанный на сочетании алгоритмов коллективного кластерного анализа и ядерных методов классификации. Идея, лежащая в его основе, состоит в рассмотре-

нии матрицы коассоциации, полученной кластерным ансамблем, как матрицы попарного сходства объектов и использовании этой матрицы как матрицы ядра (например, в методе опорных векторов *SVM* и методе ближайшего соседа с использованием ядра, *kernel kNN*). Во-первых, можно полагать, что объекты из плотной области (кластера) в пространстве признаков, с большей вероятностью имеют общие метки классов, даже если данная область имеет сложную форму. С этой точки зрения такие объекты более похожи друг на друга, чем другие точки, удаленные друг от друга на такое же расстояние, но из разных кластеров. Во-вторых, известно, что усредненная матрица коассоциации определяет полуметрику на пространстве наблюдений [9], а значит, частоты отнесения пар объектов одним и тем же кластерам можно рассматривать как показатели сходства между соответствующими точками. При этом полученная матрица зависит от выходов алгоритмов кластеризации и является менее зависимой от случайных выбросов, чем обычная матрица сходства.

Привлечение кластерного ансамбля позволяет повысить устойчивость результатов кластерного анализа в случае неопределенности в структуре данных [10,11]. В работах [12,13] были предложены различные схемы применения ансамблевой кластеризации в задаче частично контролируемой классификации. Целесообразность применения такого подхода была подтверждена экспериментальными результатами, свидетельствующими о том, что использование усредненной матрицы коассоциации в роли матрицы сходства во многих случаях существенно повышает качество решений. Это заключение было подтверждено теоретическим анализом: в [14] было доказано, что вероятность ошибки классификации уменьшается с увеличением размера ансамбля при некоторых условиях регулярности.

В данной работе предлагается новый метод частично контролируемой классификации с использованием сочетания теоретико-графового подхода, регуляризации и кластерного ансамбля. Идея предлагаемого метода состоит в использовании мало-рангового представления усредненной матрицы коассоциации для уменьшения требуемой памяти и вычислительных затрат.

Далее в работе дается математическая постановка задачи, делается краткий обзор методов частично контролируемой классификации. Описывается предлагаемая модификация и проводится ее экспериментальное исследование.

Пусть имеется генеральная совокупность объектов распознавания и конечное множество меток классов $C = \{c_1, \dots, c_k, \dots, c_K\}$. Каждый объект $a \in \Gamma$ описывается набором признаков $X = (X_1, \dots, X_d)$. Обозначим через $X_j(a)$ значение признака X_j для объекта a .

При заданных признаках набор $x(a) = (X_1(a), \dots, X_d(a))$ называется признаковым описанием (наблюдением) объекта $a \in \Gamma$.

Пусть имеется выборка $\mathbf{X} = \{x_1, \dots, x_n\}$ наблюдений за объектами $a_1, \dots, a_n$, где $x_i = x(a_i)$. В задаче частично контролируемого обучения в выборке присутствуют наблюдения двух типов:

- $\mathbf{X}_1 = \{x_1, \dots, x_{n_1}\}$ – описания объектов $a_1, \dots, a_{n_1}$ с известными метками классов $\mathbf{Y}_1 = \{y_1, \dots, y_{n_1}\}$, где $y_i \in C$ – метка класса, к которому принадлежит объект a_i , $i = 1, \dots, n_1$;

- $\mathbf{X}_0 = \{x_{n_1+1}, \dots, x_n\}$ – описания неразмеченных объектов (без ограничения общности, мы предполагаем, что первые n_1 объектов выборки являются размеченными, а следующие по порядку – неразмеченными).

Существуют два основных варианта постановки задачи. В первом варианте требуется провести индуктивное обучение, т.е. построить классификатор $f: X \rightarrow Y$, который сопоставляет номер класса как объектам из X_0 , так и произвольным новым наблюдениям. Во втором варианте постановки задачи требуется провести трансдуктивное обучение, т.е. определить метки классов $\mathbf{Y}_0 = \{y_{n_1+1}, \dots, y_n\}$ только для объектов из X_0 . Как в первом, так и во втором случае используется некоторый функционал качества классификации. В данной работе рассматривается второй вариант постановки задачи.

Предлагаемый в данной работе метод основан на комбинации коллективного кластерного анализа и теоретико-графового подхода. Для этого используется усредненная матрица коассоциации кластерного ансамбля.

Как правило, усредненная матрица коассоциации определяется при формировании кластерного ансамбля и требует памяти квадратичного размера по отношению к объему выборки [10,11]. Тем не менее, можно ослабить данное требование путем малорангового представления матрицы.

Кластерный ансамбль и малоранговая декомпозиция матрицы коассоциации

Для построения усредненной коассоциативной матрицы формируется r вариантов кластеризации объектов $\mathbf{X}$, полученных с помощью некоторого алгоритма кластерного анализа. Как правило, каждый вариант строится по выбираемым случайным образом параметрам или структурным настройкам алгоритма (таким как число кластеров, начальная инициализация, используемое подпространство признаков, тип расстояния между кластерами).

Рассмотрим варианты разбиения $\{P_l\}_{l=1}^r$, где $P_l = \{C_{l,1}, \dots, C_{l,K_l}\}$, $C_{l,k} \subset \mathbf{X}$, $C_{l,k} \cap C_{l,k'} = \emptyset$, K_l – число кластеров в l -м варианте. Для каждого варианта P_l определим матрицу $H_l = (H_l(i, j))_{i,j=1}^n$, элементы которой показывают, принадлежит ли пара x_i, x_j одному кластеру в l -м варианте или разным: $H_l(i, j) = \mathbb{I}[c_l(x_i) = c_l(x_j)]$, где $c_l(x)$ обозначает метку кластера, в который включена точка x . Взвешенная усредненная коассоциативная матрица (КМ) определена как:

$$H = (H(i, j))_{i,j=1}^n, H(i, j) = \sum_{l=1}^r w_l H_l(i, j)$$

где $w_1, \dots, w_r$ – веса элементов ансамбля, $w_l \geq 0$, $\sum_{l=1}^r w_l = 1$. Веса могут быть одинаковыми или подбираться пропорционально некоторому показателю качества каждого варианта кластеризации.

Следующее свойство позволяет значительно увеличить эффективность матричных вычислений.

Утверждение. *Взвешенная усредненная матрица ассоциации допускает малоранговое разложение в виде:*

$$H = BB^T, B = [B_1 B_2 \dots B_r], \quad (1)$$

где B - блочная матрица, $B_l = \sqrt{w_l} A_l$, A_l - матрица ассоциации для l -го разбиения, которая имеет размерность $n \times K_l$: $A_l(i, k) = \mathbb{I}[c(x_i) = k]$, $i = 1, \dots, n$, $k = 1, \dots, K_l$.

Как правило, выполняется: $m = \sum_l K_l \ll n$; таким образом, разложение (1) дает возможность сэкономить память, сохраняя разреженную B матрицу размерности $n \times m$ вместо полной усредненной матрицы коассоциации размера $n \times n$. Сложность умножения матрицы на вектор $H \cdot x$ уменьшается при этом с $O(n^2)$ до $O(nm)$.

Кластерный ансамбль и регуляризация с лапласианом графа сходства

Рассмотрим матрицу H в качестве матрицы сходства и зададим нормированный лапласиан соответствующего графа в виде: $\tilde{L} = I - \tilde{D}^{-1/2} H \tilde{D}^{-1/2}$, где $\tilde{D} = \text{diag}(\tilde{D}'_{11}, \dots, \tilde{D}'_{nn})$, $\tilde{D}'_{ii} = \sum_j H(i, j)$. Получим:

$$\tilde{D}'_{ii} = \sum_{j=1}^n \sum_{l=1}^r w_l \sum_{k=1}^{K_l} A_l(i, k) A_l(j, k) = \sum_{l=1}^r w_l \sum_{k=1}^{K_l} A_l(i, k) \sum_{j=1}^n A_l(j, k) = \sum_{l=1}^r w_l n_l(i), \quad (2)$$

где $n_l(i)$ – число точек в кластере, который включает точку x_i в l -м варианте разбиения. Подставив $\tilde{L}$ в (5), сформируем систему линейных уравнений:

$$(I_{1,0} + \beta L) F^{**} = Y_{1,0}. \quad (3)$$

Используя представление (1), эта система может быть преобразована в форму, использующую более эффективные операции с матрицами малого ранга. Для этого обозначим через $U = \tilde{D}^{-1/2} B$, тогда $\tilde{L} = I - UU^T$. Из (1) получим систему:

$$(I_{1,0} + \beta(I - UU^T)) F^{**} = Y_{1,0}. \quad (4)$$

Для ее численного решения можно использовать какой-либо из существующих приближенных итеративных алгоритмов. В данной работе используется алгоритм GDSolve [15], основанный на градиентном спуске. Сходимость алгоритма GDSolve для случая симметричной положительно определенной матрицы системы доказана в [15]. Опишем основные шаги модификации данного алгоритма GDSolveLR, использующего малоранговое представление лапласиана.

Алгоритм GDSolveLR:

Вход: $U, I_{1,0}$: разреженные матрицы в правой части системы (4);
 $Y_{1,0}$: матрица в правой части (4);
 δ - требуемая точность решения.

Выход: F^{**} : искомая матрица классификации.

Шаги алгоритма

1. $t := 0$; $F^{**}(0) := 0$;
2. **for** $k := 1 \rightarrow K$ **do**
3. $b := Y_{1,0 \cdot k}$ (k -й столбец матрицы $Y_{1,0}$);
4. **repeat**
5. вычислить невязку $r(t) := b - (I_{1,0} \cdot F^{**}_{\cdot, k}(t) + \beta F^{**}_{\cdot, k} - \beta U(U^T \cdot F^{**}_{\cdot, k}(t)))$;
6. найти оптимальную длину шага

$$\eta(t) := \frac{\mathbf{r}(t)^T \mathbf{r}(t)}{\mathbf{r}(t)^T (I_{1,0} \cdot \mathbf{r}(t) + \beta \mathbf{r}(t) - \beta U(U^T \cdot \mathbf{r}(t)))} ;$$

$$7. F_{:,k}^{**}(t+1) := F_{:,k}^{**}(t) + \eta(t) \cdot \mathbf{r}(t) ;$$

$$8. \text{until } \mathbf{r}(t) < \delta ;$$

9. end for

$$10. \text{return } F^{**}(t+1) .$$

Заметим, что на шагах 5,6 проводится умножение на матрицу лапласиана, представленную в малоранговой форме; при этом не требуется хранить в памяти полную матрицу размерности $n \times n$.

Ниже описаны основные шаги предложенного алгоритма SSC-LR-GD частично контролируемой классификации, основанного на лапласиане графа сходства, малоранговом разложении матрицы коассоциации и градиентном методе.

Алгоритм SSC-LR-GD:

Вход: $\mathbf{X}$: множество объектов с заданными классами $\mathbf{X}_1$ и неразмеченные объекты $\mathbf{X}_0$;

$\mathbf{Y}_1$: множество меток классов для размеченных объектов;

r : число запусков алгоритма кластеризации;

Ω : множество параметров работы алгоритма кластеризации; способ оценки качества группировки.

Выход: $\mathbf{Y}_0$: предсказанные метки классов для объектов из $\mathbf{X}_0$.

Шаги алгоритма

1. Сформировать r вариантов кластеризации объектов из $\mathbf{X}$, случайным образом выбирая из Ω параметры алгоритма кластерного анализа; найти веса $w_1, \dots, w_r$ вариантов (оценки качества группировки);

2. Вычислить матрицу нормализованного лапласиана в малоранговом представлении.

3. Найти классификационную матрицу F^{**} с помощью алгоритма GDSolveLR;

4. Определить метки для $\mathbf{Y}_0$, используя матрицу $F = F^{**}$.

end.

В компьютерной реализации данного алгоритма используется метод К-средних в качестве базового алгоритма кластеризации, что позволяет достичь линейной трудоемкости по отношению к размеру входных данных.

С разработанным алгоритмом были проведены численные эксперименты, цель которых заключалась в проверке его эффективности в условиях различного объема данных и наличии шумовых эффектов. Качество работы алгоритма определялось на а) многократно сгенерированных искусственных данных, подчиняющихся заданному распределению и б) реальной задаче распознавания гиперспектрального изображения.

Смесь нормальных распределений

В данном примере рассмотрены наборы данных, сгенерированные из смеси пяти многомерных нормальных распределений $\mathcal{N}(a_i, \sigma_x I)$ при равных весах; $a_i \in \mathbf{R}^d$, $i = 1, \dots, 5$, $d = 8$; величина σ_x является заданным параметром. Для исследования устой-

чивости работы алгоритма в условиях шума дополнительно генерируются две независимые случайные величины, подчиняющиеся равномерному распределению $\mathcal{U}(0, \sigma_z)$, которые используются в качестве дополнительных «зашумленных» признаков. Параметр шума $\sigma_z = 5$. На рисунке 1 приведен пример сгенерированных данных. При моделировании многократно генерировались выборки объема n в соответствии с заданной моделью смеси распределений. В эксперименте 10% точек, выбранных случайным образом для каждой компоненты, составляли размеченную часть; остальные точки были включены в неразмеченную подвыборку.

Варианты решений кластерного ансамбля формировались путем случайного выбора точек в качестве начальных центроидов в алгоритме К-средних (количество кластеров равно 10). Число элементов ансамбля равно: $r = 10$. Веса

Рисунок 1 – Сгенерированные данные
(проекция на плоскость X_1, X_2): $n = 1000$, $\sigma_x = 1$

элементов ансамбля одинаковы: $w_i = 1/r$. Параметр регуляризации β оценивался с использованием кроссвалидации и поиска по сетке в интервале $[0.001, 0.5]$ с шагом 0.005; наилучшие результаты были получены для значения $\beta = 0.1$. Параметр $\delta = 10^{-5}$.

С целью сравнения был рассмотрен алгоритм (обозначенный как SSC-RBF), который использовал стандартную матрицу подобия, оцененную с помощью RBF ядра (параметр $\sigma = 4$).

Для повышения степени достоверности результатов оценки качества были усреднены по 40 экспериментам. Анализ статистической значимости различий между оценками проводился с помощью парного t -критерия Стьюдента.

В таблице 1 представлены результаты экспериментов. В дополнение к усредненным показателям точности решений, в таблице приведены усредненные времена работы алгоритмов. Для алгоритма SSC-LR-GD отдельно указано время генерации ансамбля и время, потраченное на матричные преобразования (в секундах). Жирным шрифтом выделены оценки точности, статистически значимо превышающие оценки точности для сравниваемого алгоритма (p -value $< 10^{-5}$).

Таблица 1 – Результаты экспериментов со смесью распределений при различных объемах выборки и значениях параметра .

n	σ_x	SSC-LR-GD			SSC-RBF	
		точность	t_{ens} (sec)	t_{matr} (sec)	точность	время (sec)
1000	1	1.000	0.06	0.10	1.000	0.32
	3	0.985	0.07	0.02	0.982	0.32
	5	0.874	0.13	0.11	0.817	0.35
3000	1	1.000	0.10	0.48	1.000	5.42
	3	0.986	0.13	0.10	0.984	5.29
	5	0.878	0.23	0.48	0.848	5.48
10^5	1	1.000	2.05	25.69	-	-
10^6	1	1.000	49	443	-	-

Результаты показывают, что в данных экспериментах алгоритм SSC-LR-GD показал сравнимую или даже значимо большую точность, чем SSC-RBF. Для большого объема данных ($n = 10^5$, $n = 10^6$) алгоритм SSC-RBF отказался работать из-за недопустимых требований к памяти (74.5 ГБ и 7450.6 ГБ). Основное время работы алгоритма SSC-LR-GD пришлось на этап итеративных вычислений. При этом время работы алгоритма SSC-LR-GD оказалось гораздо меньше, чем у аналогичного алгоритма SSC-RBF, не использующего малоранговое представление матрицы сходства.

Анализ гиперспектрального изображения

Для экспериментального исследования разработанного алгоритма было использовано гиперспектральное изображение Indian Pines [16] размером 145 на 145 пикселей, которое содержит 224 спектральных канала в диапазоне 400-2500 nm. На рисунке 2(а) показан RGB-композит изображения, а на рисунке 2(б) приведено эталонное разбиение изображения на 16 тематических классов. На снимке имеются неразмеченные пиксели, которые не отнесены ни к одному из классов. Данные пиксели были исключены из рассмотрения при анализе.

Рисунок 2 – Гиперспектральное изображение Indian Pines: RGB композит, (b) размеченные данные.

В эксперименте размеченная часть составила 1% точек, выбранных случайным образом для каждой компоненты. Чтобы уменьшить эффект коррелированности спектральных каналов, предварительно из исходных данных с помощью метода главных компонент были сформированы 10 новых признаков. Сформированная таблица данных подавалась на вход алгоритма SSC-LR-GD. Размер ансамбля был равен $r = 10$, а различные варианты разбиения получались варьированием числа кластеров в интервале $[1000, 1000 + r]$. Остальные параметры совпадали с описанными в предыдущем эксперименте.

Проводилось сравнение с методом опорных векторов SVM, использующим RBF ядро (выбирались параметры алгоритма, рекомендуемые по умолчанию).

Для повышения статистической надежности результатов, на основе случайно отобранных размеченных подвыборок было получено 20 вариантов решений для каждого из алгоритмов. При этом вычислялись усредненные значения точности классификации неразмеченных пикселей изображения, метки которых не использовались при обучении. Среднее время работы алгоритма SSC-LR-GD составило около 1.5 мин, алгоритма SVM – 0.3 мин. В итоге средняя точность SSC-LR-GD составила 0.557, а SVM – 0.513. Парный t -тест показал значимое повышение точности у алгоритма SSC-LR-GD (p -value $< 10^{-5}$).

В работе рассмотрена задача распознавания образов в случае неполной обучающей информации. С использованием комбинации методик, основанных на регуляризации лапласиана графа сходства, коллективном кластерном анализе и малоранговой матричной декомпозиции, разработан алгоритм решения данной задачи.

Проведено экспериментальное исследование предложенного алгоритма с помощью статистического моделирования. В проведенных экспериментах алгоритм показал сравнимую или даже значимо большую точность, чем аналогичный алгоритм, не использующий ансамблевый кластерный анализ и малоранговое представление матрицы коассоциации. При этом в случае большого объема данных, стандартный алгоритм отказался работать из-за недопустимых требований к памяти. Среднее время работы предложенного алгоритма оказалось намного меньше, чем у аналогичного стандартного.

Благодарности. Работа проведена при поддержке Российского фонда фундаментальных исследований (гранты 18-07-00600а, 18-29-0904мк).

ЛИТЕРАТУРА

- 1 Zhu X. Semi-supervised learning literature survey//Tech. Rep. (Department of Computer Science, Univ. of Wisconsin, Madison), 2008. P.1530.
- 2 Загоруйко Н.Г. Методы распознавания и их применение//Советское радио. – 1972. – 182с. [Zagorujko N.G. Metody raspoznavaniya i ih primeneniye//Sovetskoe radio. – 1972. – 182s.]
- 3 Camps-Valls G., Marsheva T., Zhou D. Semi-supervised graph-based hyperspectral image classification // IEEE Transactions on Geoscience and Remote Sensing. 2007. P.3044-3054.
- 4 Wu M., Scholkopf B. Transductive Classification via Local Learning Regularization// Artificial Intelligence and Statistics. 2007. P. 628-635.
- 5 Zhou D., Bousquet O., Lal T., Weston J., Scholkopf B. Learning with local and global consistency // In Advances in Neural Information Processing Systems. 2003. V.16. P.321-328.

6 Belkin M., Niyogi P., Sindhvani V. Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples // J. Mach. Learn. Res. 2006. V.7. - P.2399-2432.

7 Berikov V.B., Pestunov I.A., Karaev N.M., Tewari A., "Recognition of hyperspectral images with use of cluster ensemble and semisupervised learning", CEUR Workshop Proceedings. Vol.2033,2017.P.60-64. http://ceur-ws.org/Vol-2033/10_paper.pdf

8 Berikov V., Cherikbayeva L., "Searching for Optimal Classifier Using a Combination of Cluster Ensemble and Kernel Method", CEUR Workshop Proceedings. – Vol. 2098. – 2017. – P. 45-60. <http://ceur-ws.org/Vol-2098/paper4.pdf>

9 Berikov V.B. Construction of an optimal collective decision in cluster analysis on the basis of an averaged co-association matrix and cluster validity indices // Pattern Recognition and Image Analysis. 2017. P.153-165.

10 Berikov V.B. Weighted ensemble of algorithms for complex data clustering // Pattern Recognition Letters, 2014. P. 99-106.

11 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.

12 Yu G. X., Feng L., Yao G. J., Wang, J. Semi-supervised classification using multiple clusterings // Pattern Recognition and Image Analysis. 2017. Vol. 4. P. 681-668.

13 Berikov V., Karaev N., Tewari A. Semi-Supervised Classification with Cluster Ensemble // Proceedings of 2017 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON). 18-22 Sep 2017, Novosibirsk Akademgorodok, Russia. – P. 245-250.

14 Amirgaliyev Y., Berikov V., Latuta K., Bekturgan K., Cherikbayeva L. Group Approach to Solving the Tasks of Recognition // Yugoslav Journal of Operations Research. 2019. V.1. P.212.

15 Vishnoi, Nisheeth K. Lx= b Laplacian Solvers and Their Algorithmic Applications // Foundations and Trends in Theoretical Computer Science 8.1–2. 2013. P. 1-141.

16 Hyperspectral Remote Sensing Scenes, http://www.ehu.eus/ccwintco/index.php_title=Hyperspectral_Remote_Sensing_Scenes (дата обращения: 17.04.2019).

**В.Б. БЕРИКОВ¹, Е.Н. АМИРГАЛИЕВ², Л.Ш. ЧЕРИКБАЕВА³,
Б.А. ТУЛЕГЕНОВА⁴**

¹С.Л. Соболев атындағы математика институты, Ресей

²ҚР БҒМ ҒК Ақпараттық-есептеу техникасы институты, Қазақстан

³Әл-Фараби атындағы Қазақ ұлттық университеті, Қазақстан

⁴Қ.И. Сәтбаев атындағы Қазақ ұлттық техникалық зерттеу университеті,
Қазақстан

E-mail:berikov@math.nsc.ru;amir_ed@mail.ru;lyalya_sh@mail.ru;
tulegenova_bakhit@mail.ru

ОҚЫТУ АҚПАРАТТАРЫ ТОЛЫҚ ЕМЕС ЖАҒДАЙЫНДА КЛАСТЕРЛІК АНСАМБЛЬ ЖӘНЕ АЗ РАНГІЛІ МАТРИЦАЛЫҚ ДЕКОМПОЗИЦИЯНЫ ҚОЛДАНА ОТЫРЫП КЛАССИФИКАЦИЯЛАУ

Бұл жұмыста оқыту ақпараттары толық емес жағдайдағы бейнені тану есебі шешіледі. Ұқсастық графының лапласианын реттеу мен топтық кластерлік талдау сәйкестендірілетін шешу әдісі ұсынылады. Жадыны үнемдеп және есептеу жылдамдығын жоғарылатуға мүмкіндік беретін кластерлік талдаудың коассоциациялық матрицасын азрангілі түрде ұсыну әдісі

қолданылады. Тесттік есептер мен нақты гиперспектральды бейнелердегі сандық эксперименттер шулы деректердің болған жағдайда да ұсынылып отырған әдістің тиімді екендігін көрсетеді,

Түйін сөздер: тану, классификация, ансамбль, коассоциативті матрица, топтық кластердік талдау.

**V.B. BERIKOV, E.N. AMIRGALIYEV, L. SH. CHERIKBAYEVA,
B.A. TULEGENOVA**

¹*S.L. Sobolev Institute of Mathematics SB RAS, Novosibirsk State University, Russia*

²*Institute of Information and Computing Technologies KN MES RK, Kazakhstan*

³*Al-Farabi Kazakh National University, Kazakhstan*

⁴*Satbayev University, Kazakhstan*

E-mail: berikov@math.nsc.ru; amir_ed@mail.ru; lyalya_sh@mail.ru;

tulegenova_bakhit@mail.ru

CLASSIFICATION WITH INCOMPLETE TRAINING INFORMATION USING A CLUSTER ENSEMBLE AND LOW-RANK MATRIX DECOMPOSITION.

The paper is devoted to solve the pattern recognition problem with incomplete learning data. The solution method, which combines similarity graph with Laplacian Regularization and collective clustering is proposed. The low-rank decomposition of co-association matrix for cluster ensemble is used, which allows to speed up the computations and keep memory. Experimental results on test tasks and on real hyperspectral image demonstrate the effectiveness of proposed method, including with noisy data.

Key words: recognition, classification, ensemble, co-association matrix, collective clustering.