КЛАСТЕРНЫЙ АНАЛИЗ ДАННЫХ НА ОСНОВЕ СВЯЗАМИ ОБЪЕКТОВ ПО ЛОКАЛЬНЫМИ ОБЪЕКТАМИ

Игнатьев Н.А
Национальный университет Узбекистана имени Мирзо
Улугбека n_ignatev@rambler.ru
Национальный университет Узбекистана имени Мирзо Улугбека
Рамазонов Ш.Ш
ramazonov377@gmail.com

Аннотация: В задачах анализа данных и машинного обучения важную роль играет структура обучающей выборки. Большие выборки могут содержать избыточные или повторяющиеся объекты, что увеличивает вычислительную сложность и может снижать эффективность алгоритмов классификации. В данной работе проводится анализ устойчивости выборок, полученных с помощью алгоритма селекции объектов. Для оценки структурных свойств используется новый номинальный признак, формируемый на основе состава *k* ближайших соседей. Устойчивость данного признака оценивается с помощью показателя. Для различных уровней селекции формируется матрица устойчивости, после чего проводится корреляционный анализ между выборками. Кроме того, применяется метод главных компонент (PCA) для визуального анализа структурных различий между выборками.

Ключевые слова: селекция выборки, *k* ближайших соседей, устойчивость признаков, корреляционный анализ, PCA.

DATA CLUSTER ANALYSIS BASED ON OBJECT RELATIONS BY LOCAL OBJECTS

Ignatev N.A
National University of Uzbekistan named after Mirzo Ulugbek **Sh.Sh. Ramazonov**
n_ignatev@rambler.ru
National University of Uzbekistan named after Mirzo Ulugbek
Sh.Sh. Ramazonov
ramazonov377@gmail.com

Abstract: In data analysis and machine learning tasks, the structure of the training dataset plays an important role. Large datasets may contain redundant or duplicated objects, which increases computational complexity and may reduce the efficiency of classification algorithms.In this paper, the stability of datasets obtained using an object selection algorithm is analyzed. To evaluate the structural properties of the data, a new nominal feature is introduced based on the composition of the ***k*-nearest neighbors**. The stability of this feature is evaluated using a specific indicator.For different levels of selection, a stability matrix is constructed, after which a correlation analysis between the datasets is performed. In addition, **Principal Component Analysis (PCA)** is applied for the visual analysis of structural differences between the datasets.

Keywords: sample selection, k-nearest neighbors, feature stability, correlation analysis, PCA.

BERILGANLARNI LOKAL QO‘SHNILAR ASOSIDA OBYEKTLAR BOG‘LANISHLARI ORQALI KLASTER TAHLILI

Ignatev N.A
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy Universiteti n_ignatev@rambler.ru
Sh.Sh. Ramazonov
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy Universiteti ramazonov377@gmail.com

Annotatsiya: Ma’lumotlarni tahlil qilish va mashinaviy o‘rganish masalalarida o‘rgatuvchi tanlanmaning tuzilishi muhim ahamiyatga ega. Katta hajmdagi tanlanmalar ortiqcha yoki takrorlanuvchi obyektlarni o‘z ichiga olishi mumkin, bu esa hisoblash murakkabligini oshiradi hamda klassifikatsiya algoritmlarining samaradorligini kamaytirishi mumkin.Ushbu ishda obyektlarni seleksiya qilish algoritmi yordamida hosil qilingan tanlanmalarning barqarorligi tahlil qilinadi. Ma’lumotlarning strukturaviy xususiyatlarini baholash uchun **k ta eng yaqin qo‘shni tarkibiga asoslangan yangi nominal belgi** kiritiladi. Ushbu belgining barqarorligi maxsus ko‘rsatkich yordamida baholanadi.Seleksiya darajalarining turli qiymatlari uchun barqarorlik matritsasi shakllantiriladi va tanlanmalar o‘rtasida korrelyatsion tahlil amalga oshiriladi. Bundan tashqari, tanlanmalar o‘rtasidagi strukturaviy farqlarni vizual tahlil qilish uchun **asosiy komponentlar usuli (PCA)** qo‘llaniladi.

Kalit so‘zlar: tanlanma seleksiyasi, k ta eng yaqin qo‘shnilar, belgilar barqarorligi, korrelyatsion tahlil, PCA.

Введение. В задачах анализа данных, машинного обучения и распознавания образов качество обучающей выборки оказывает существенное влияние на эффективность построенных моделей. Результаты классификации или регрессии во многом зависят от структуры обучающего множества, его объёма и распределения объектов в пространстве признаков. Поэтому формирование обучающих выборок, определение их оптимального объема и анализ структурных различий между различными выборками являются важными направлениями исследований в области машинного обучения.

Одним из наиболее распространённых подходов для оценки обобщающей способности моделей является кросс-валидация (cross-validation)**[1]**. В данном методе исходный набор данных разбивается на несколько частей, и модель обучается и тестируется на различных комбинациях этих частей. Наиболее популярным вариантом является *k*-fold cross-validation, при котором выборка делится на *k* равных частей[2], и каждая из них поочередно используется в качестве тестовой. Такой подход позволяет более объективно оценить качество модели и выявить влияние структуры обучающей выборки на результаты обучения.

Наряду с кросс-валидацией в машинном обучении широко применяются ансамблевые методы (ensemble learning)[3]. Основная идея ансамблевых алгоритмов заключается в объединении нескольких моделей для получения более устойчивого и точного результата. К наиболее известным ансамблевым методам относятся bagging, boosting и random forest[4]. В методе bagging из исходной выборки формируются случайные подвыборки, на основе которых строятся независимые модели. В алгоритмах boosting модели обучаются последовательно, и каждая новая модель стремится исправить ошибки предыдущей. Такие подходы показывают, что структура обучающей выборки и способ её формирования существенно влияют на качество итоговой модели.

Существует несколько способов формирования обучающих выборок[5]. Наиболее простым является случайная выборка (random sampling)**[6]**, при которой исходный набор данных случайным образом разделяется на обучающую и тестовую части. Другим распространённым методом является стратифицированная выборка (stratified sampling), позволяющая сохранить пропорциональное

распределение классов. Кроме того, активно исследуются методы селекции объектов и отбора прототипов**[7]**, направленные на выделение наиболее информативных элементов обучающей выборки. Такие методы позволяют уменьшить размер обучающего множества, сохраняя при этом его основные структурные свойства.

Сравнение различных обучающих выборок также является важной задачей анализа данных[8]. Для этого используются различные статистические и геометрические методы. В частности, степень сходства между выборками может оцениваться с помощью коэффициентов корреляции[9], различных метрик расстояния и методов кластерного анализа. Кроме того, для визуализации высокоразмерных данных широко применяется метод главных компонент (PCA)**[10]**, позволяющий представить данные в пространстве меньшей размерности и выявить структурные различия между выборками.

В данной работе проводится анализ структурной устойчивости выборок, полученных с помощью алгоритма селекции объектов. Сначала формируются несколько редуцированных обучающих выборок с различными уровнями селекции. Далее для каждой выборки формируется новый синтетический признак на основе состава k ближайших соседей, отражающий локальную структуру данных. Значением признака является число объектов своего классе среди $\boldsymbol{k}$ ближайших. Множества допустимых значений от о до $\boldsymbol{k}$.

Средством для анализа служит устойчивость признака со множеством допустимых значении от 0.5 до 1.

Проверяется наличие закономерностей на каждой обучающей выборке и на их совокупности.

Устойчивость данного признака оценивается с использованием показателя.

Для различных уровней селекции формируется матрица устойчивости, после чего проводится анализ взаимосвязей между выборками с использованием коэффициентов корреляции Пирсона и Спирмена. Дополнительно применяется метод главных компонент (PCA) для визуального анализа структурных различий между редуцированными выборками. Такой подход позволяет исследовать влияние уровня селекции на структурные свойства обучающего множества и на локальные характеристики данных.

Многие существующие методы селекции направлены на уменьшение размера обучающей выборки. Однако в большинстве исследований недостаточно анализируется сохранение структурных свойств данных после редукции. Предлагаемый подход отличается тем, что проводится анализ устойчивости выборок с использованием нового номинального признака, сформированного на основе состава k ближайших соседей. Кроме того, проводится корреляционный анализ между различными редуцированными выборками и используется метод главных компонент для визуализации различий между ними.

2.Постановка задачи

Пусть дана выборка: $E_0=\{S_1,\ldots,S_m\}$, Каждый объект описывается вектором признаков: $X(n)=(x_1,\ldots,x_n)$.Выборка разделена на два класса: K_1 и K_2

Требуется:

- Сформировать наборы обучающих и валидационных выборок.
- Построить синтетические признаки по $\boldsymbol{k}$ ближайшим соседям.
- Произвести корреляционный анализ по значением устойчивости

Алгоритм 1:

- Выборка $E_0=\{S_1,\ldots,S_m\}$, где каждый объект описан вектором разнотипных признаков $X(n)=(x_1,\ldots,x_n)$.
- Выборка E_0 разделена на два непересекающихся класса: K_1 и K_2.
- Задана функция расстояния (метрика) $\rho(S_i, S_j)$.

Определим вспомогательные функции для каждого объекта $S_i \in E_0$:

- $r(S_i)$ — расстояние от объекта S_i до ближайшего к нему объекта противоположного класса (радиус до ближайшего "врага").

• $g(S_i)$ — количество объектов своего класса, находящихся строго внутри гипершара с центром в S_i и радиусом $r(S_i)$:

• $g(S_i)=\{S_j \in K_t | \rho(S_i,S_j)<r(S_i)\}$, где K_t — класс объекта S_i.

• $z(S_i,k)$ — расстояние от объекта S_i до его k-го ближайшего соседа.

• $d(S_i)= |z(S_i,k)-r(Si)|/(z(S_i,k)+r(S_i))$.

Цель алгоритма: построить сокращённое (селектированное) подмножество исходной выборки.

Шаги алгоритма:

1. Инициализация.

Вход: множество объектов E_0. Для всех $S_i \in E_0$ вычислить $r(S_i)$, $g(S_i)$, $z(S_i,k)$, $d(S_i)$. Инициализировать рабочий набор $D=E_0$ (копия исходной выборки). Определить множество удаляемых объектов $R=\emptyset$.

2. Формирование приоритетной очереди.

Упорядочить объекты из множества $\{S_i \in D | d(S_i) \geq 0\}$ по возрастанию значения $d(S_i)$. Объекты с наименьшим $d(S_i)$ получают наивысший приоритет. Если для всех объектов в D выполняется $d(S_i)<0$, перейти к шагу 6.

3. Выбор кандидата на удаление.

Взять объект S_i с наименьшим значением $d(S_i)$ (первый в упорядоченном списке).

4. Определение окрестности и проверка условия.

Пусть S_i принадлежит классу K_t. Построить окрестность объекта S_i:

$L=\{S_j \in (D \cap K_t) | \rho(S_j,S_i)<r(S_j)$ и $g(S_j)>1\}$.

Это множество других объектов того же класса из D, для которых S_i лежит внутри их собственной сферы влияния (радиуса $r(S_j)$) и у которых текущее количество соседей внутри этой сферы больше единицы. «Заблокировать» объект S_i, присвоив $d(S_i)=-d(Si)$ (это предотвратит его повторный выбор на шаге 3).

Если $|L|=1$, то удалить S_i из рабочего множества:

$D=D \setminus \{S_i\}, R=R \cup \{S_i\}$.

5. Коррекция характеристик соседей (при $|L|>1$).

Для каждого объекта $S_j \in L$:

○уменьшить счётчик соседей: $g(S_j)=g(S_j)-1$;

○если после уменьшения $g(S_j)=1$ (т.е. S_j остался единственным объектом в своей $r(S_j)$- окрестности), удалить S_j из рабочего множества: $D=D \setminus \{S_j\}, R=R \cup \{S_j\}$. Перейти к шагу 2.

Алгоритм 2:

• Выборка $E_0=\{S_1,\ldots,S_m\}$, каждый объект которой описан вектором разнотипных признаков $X(n)=(x_1,\ldots,x_n)$. Выборка E_0 разделена на два непересекающихся класса: K_1 и K_2.

• Задана функция расстояния (метрика) $\rho(S_i,S_j)$.

Обозначим через $O(S,k)$ множество возможных соотношений числа представителей классов K_1 и K_2 среди k ближайших соседей объекта S. Это множество имеет вид

$$O(S,k) = \{(i, k-i) \mid i = 0,1, \ldots, k\},$$

где i — количество соседей из класса K_1, а $k-i$ — из класса K_2. Всего существует $k+1$ различных вариантов.

Поставим в соответствие каждому такому варианту число $\mu \in \{1,2,\ldots,k+1\}$, которое будем считать градацией нового номинального признака, характеризующего объекты исходной выборки.

Значение функции принадлежности для градации μ и значение устойчивости признака вычисляются соответственно по формулам:

$$f_k(\mu) = \frac{d_{1k}(\mu)/|K_1|}{d_{1k}(\mu)/|K_1| + d_{2k}(\mu)/|K_2|},$$

и

$$g_k = \frac{1}{h}\sum_{\mu=1}^{k+1}\begin{cases} n(\mu)\cdot f_k(\mu), & f_k(\mu) \geq 0.5, \\ n(\mu)\cdot(1-f_k(\mu)), & f_k(\mu) < 0.5, \end{cases}$$

где $d_{1k}(\mu)$ и $d_{2k}(\mu)$ — число объектов класса K_1 и K_2 соответственно, имеющих градацию μ; $n(\mu)=d_{1k}(\mu)+d_{2k}(\mu)$.

Для каждого объекта определяется k ближайших соседей. Соотношение классов среди этих соседей записывается как (i, k-i), где i – число соседей из класса K_1. На основе этих комбинаций вводится новый номинальный признак: $\mu \in \{1,2,...,k+1\}$Для каждого значения μ вычисляются: $d_{1k}(\mu)$ – число объектов класса K_1 $d_{2k}(\mu)$– число объектов класса $K_2 n(\mu)=d_{1k}(\mu)+d_{2k}(\mu)$. Устойчивость признака определяется по формуле:

$$g_k(\mu) = |d_{1k}(\mu) - d_{2k}(\mu)| / n(\mu)$$

3. Экспериментальные результаты

Эксперимент на данных GERMAN[2] в ходе эксперимента были сформированы несколько редуцированных выборок с параметрами селекции: K = {7, 32, 57, 82, 107} Для каждой выборки проводился анализ устойчивости при различных значениях параметра k от 3 до 32.

Таблица 1 – Значения устойчивости признаков для различных уровней селекции объектов

k	K=7	K=32	K=57	K=82	K=107
3	0,775281	0,776062	0,780876	0,768924	0,781609
4	0,782772	0,787645	0,776892	0,788845	0,793103
5	0,794007	0,803089	0,776892	0,792829	0,808429
6	0,794007	0,795367	0,772908	0,780876	0,793103
7	0,794007	0,791506	0,780876	0,772908	0,796935
8	0,786517	0,787645	0,772908	0,76494	0,785441
9	0,76779	0,776062	0,780876	0,776892	0,796935
10	0,775281	0,783784	0,796813	0,784861	0,800766
11	0,779026	0,779923	0,792829	0,780876	0,804598
12	0,790262	0,779923	0,788845	0,780876	0,789272
13	0,801498	0,76834	0,796813	0,784861	0,793103
14	0,794007	0,772201	0,788845	0,780876	0,789272
15	0,801498	0,776062	0,788845	0,776892	0,777778
16	0,808989	0,779923	0,788845	0,796813	0,789272
17	0,808989	0,791506	0,792829	0,776892	0,789272
18	0,805243	0,787645	0,796813	0,800797	0,800766
19	0,797753	0,799228	0,784861	0,800797	0,804598
20	0,790262	0,803089	0,792829	0,796813	0,793103
21	0,779026	0,803089	0,792829	0,796813	0,800766
22	0,786517	0,803089	0,804781	0,796813	0,804598
23	0,786517	0,80695	0,800797	0,808765	0,800766
24	0,797753	0,80695	0,808765	0,812749	0,804598
25	0,790262	0,818533	0,808765	0,804781	0,816092
26	0,782772	0,795367	0,800797	0,796813	0,804598
27	0,805243	0,795367	0,808765	0,804781	0,804598
28	0,790262	0,80695	0,828685	0,812749	0,800766
29	0,782772	0,799228	0,812749	0,804781	0,800766
30	0,790262	0,80695	0,796813	0,812749	0,800766
31	0,790262	0,795367	0,796813	0,796813	0,800766

32	0,782772	0,783784	0,796813	0,804781	0,800766

В результате была сформирована матрица устойчивости. Корреляционный анализ показал, что выборки с параметрами *K*=32, *K*=57, *K*=82 и *K*=107 демонстрируют умеренную и сильную взаимосвязь, тогда как выборка с K=7 практически не коррелирует с остальными.

Таблица 2 – Коэффициенты корреляции между сформированными обучающими выборками

	K=7	**K=32**	**K=57**	**K=82**	**K=107**
K=7	1	0,030458	0,094427	0,154737	-0,12315
K=32	0,030458	1	0,443935	0,680111	0,666297
K=57	0,094427	0,443935	1	0,72345	0,480784
K=82	0,154737	0,680111	0,72345	1	0,66651
K=107	-0,12315	0,666297	0,480784	0,66651	1

Метод главных компонент (PCA) подтвердил эти результаты. В пространстве главных компонент выборка K=7 оказалась значительно удалённой от остальных, что свидетельствует о существенном изменении структуры данных при сильной редукции.

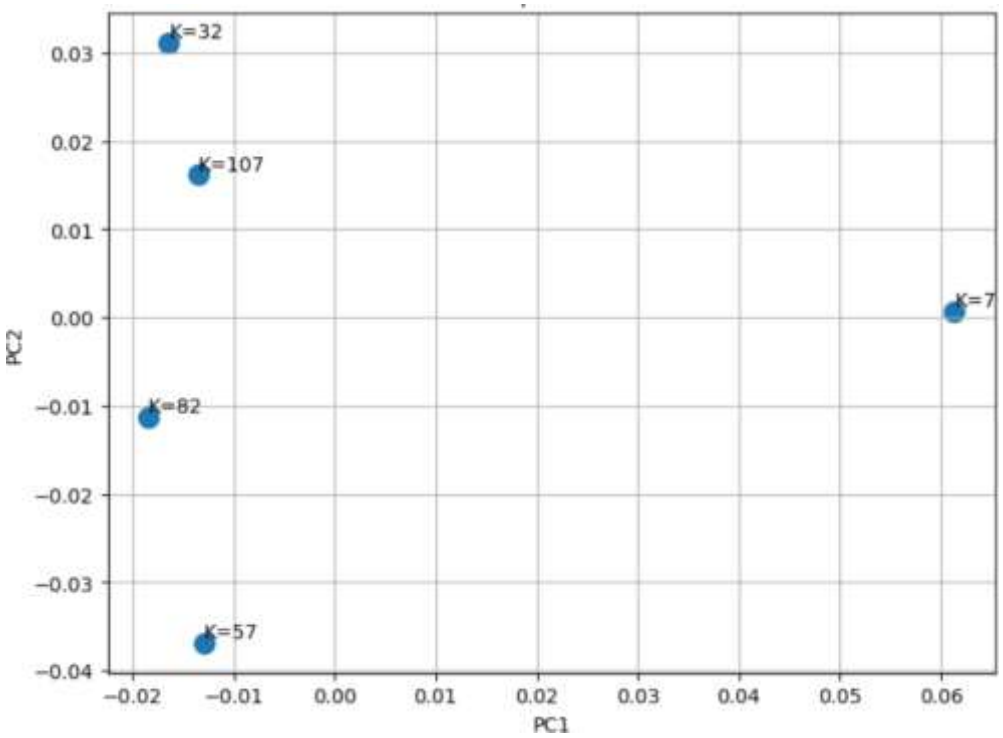

Рисунок 1–Структура локальных соседей объектов

Заключение. В данной работе был предложен подход к анализу структурной устойчивости обучающих выборок, основанный на использовании информации о составе *k* ближайших соседей объектов.Основной вклад исследования заключается в формировании синтетического номинального признака, отражающего локальное распределение классов в окрестности объектов. На основе данного признака вводится показатель устойчивости, позволяющий количественно оценивать изменения локальной структуры данных при различных уровнях редукции обучающей выборки. Для проверки предложенного подхода были проведены вычислительные эксперименты на наборе данных GERMAN Credit Data[2]. Полученные результаты показали, что при умеренных уровнях селекции структурные свойства обучающих выборок сохраняются в значительной степени, тогда как при сильной редукции наблюдаются существенные изменения структуры данных.Проведённый корреляционный анализ и визуализация данных с использованием метода главных компонент подтвердили эффективность предложенного подхода для анализа структурных характеристик редуцированных выборок. Предложенный метод может быть использован при разработке алгоритмов селекции объектов и анализе обучающих данных в задачах машинного обучения и

интеллектуального анализа данных. В дальнейшем планируется расширить экспериментальные исследования за счёт использования других наборов данных и проведения сравнительного анализа с существующими методами селекции объектов.

Список литературы

1. Игнатьев Н.А., Акбаров Б.Х. Оценка близости структур отношений объектов обучающей выборки на многообразиях наборов латентных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2023. № 65. С. 69–78.

2. *https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data*

3. Н. А. Игнатьев Выбор целевых признаков для классификации и кластерного анализа структур отношений объектов// InternationalJournalof Open Information Technologies ISSN: 2307-8162 vol. 14, no. 3, 2026

4. Hart P. The condensed nearest neighbor rule // IEEE Transactions on Information Theory. 1968. Vol. 14, No. 3. P. 515–516. https://doi.org/10.1109/TIT.1968.1054155

5. Wilson D. Asymptotic properties of nearest neighbor rules using edited data // IEEE Transactions on Systems, Man, and Cybernetics. 1972. Vol. 2, No. 3. P. 408–421. https://doi.org/10.1109/TSMC.1972.4309137

6. Tomek I. Two modifications of CNN // IEEE Transactions on Systems, Man, and Cybernetics. 1976. Vol. 6. P. 769–772.

7. Aha D., Kibler D., Albert M. Instance-based learning algorithms // Machine Learning. 1991. Vol. 6. P. 37–66. https://doi.org/10.1023/A:1022689900470

8. Wilson D., Martinez T. Reduction techniques for instance-based learning algorithms // Machine Learning. 2000. Vol. 38. P. 257–286. https://doi.org/10.1023/A:1007626913721

9. García S., Derrac J., Cano J., Herrera F. Prototype selection for nearest neighbor classification: taxonomy and empirical study // IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012. Vol. 34. No. 3. P. 417–435. https://doi.org/10.1109/TPAMI.2011.142

KVANT KALITLARINI TA’MINLASH USULI ORQALI XAVFSIZ AXBOROT ALMASHINUVI

[1] D.T. Muhamediyeva, [1] Tagaev Farxad Abduvaxabovich
«Toshkent irrigatsiya va qishloq xo’jaligini mexanizatsiyalash muhandislari instituti» milliy tadqiqot universiteti
dilnoz134@rambler.ru

Annotatsiya: Ushbu maqolada kvant kriptografiyasining BB84 protokoli asosida xavfsiz kalit almashinuvi modeli ishlab chiqildi va simulyatsiya qilindi. Tadqiqotda Alice va Bob orasidagi kvant kanal orqali uzatilgan fotonlar kvant holatlarida kodlanib, ularni o‘qishda bazalarning tasodifiy tanlanishi orqali xavfsizlik ta’minlandi. Eavesdropping (Eve) aralashuvi aniqlanishi uchun xatolik ko‘rsatkichi hisoblandi. Kalitni generatsiya qilgandan so‘ng, ma’lumotlarni himoyalangan tarzda uzatish uchun XOR asosidagi sodda ECC shifrlash algoritmi qo‘llandi. Natijalar BB84 protokoli orqali shakllangan kalit yordamida axborot maxfiyligini samarali saqlash mumkinligini ko‘rsatdi.

Kalit so‘zlar. BB84 protokoli, kvant kalit taqsimoti (QKD), kvant kriptografiyasi, XOR shifrlash, Aer simulyatori.