УДК 004.942

НЕКОТОРЫЕ МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ ДЛЯ ВЕРТИКАЛЬНОГО РАЗРЕЗА СКОРОСТИ ЗВУКА В ОКЕАНЕ

В.О. Захаров

В статье приводится сравнение методов снижения размерности применительно к профилям скорости звука в морских волноводах. Приводятся методы, основанные на машинном обучении. Производится сравнение методов и выбор наиболее подходящего метода для решения данной задачи.

The article presents methods for comparing sound sizes as applied to the sound velocity profile in sea waveguides. The methods based on machine learning are given. Performs a comparison of methods and selection of the most appropriate parameters to solve this problem.

КЛЮЧЕВЫЕ СЛОВА
Нейронная сеть, метод главных компонент, скорость звука, разреженное моделирование.

ДЛЯ ЦИТАТЫ
В.О. Захаров. Некоторые методы снижения размерности для вертикального разреза скорости звука в океане // Моделирование и анализ данных. 2019. №2. С.48-56.

V.O. Zakharov. Some methods of reducing the dimension for a vertical cut of the speed of sound in the ocean. Modelirovaniye i analiz dannykh=Modelling and data analysis *(Russia)*. 2019, no.2, pp.48-56.

1. ВВЕДЕНИЕ

Для расчета скорости звука в океанической среде используется формула Вильсона [6], предложенная им в 1960 году. График (см.рис.1.1), показывающий зависимость скорости звука от глубины, будем называть вертикальным распределением скорости звука (ВРСЗ). Так как для формулы Вильсона нужно очень много информации об океанической среде, в основном используются приближенные модели, например, профиль Munka – идеализированный профиль скорости звука. Такой профиль считается как среднее значение скорости в океанической среде – 1500 м/с умноженное на гладкую функцию [6]. Этот подход имеет малое отношение к реальности, так как результат получается очень усредненным.

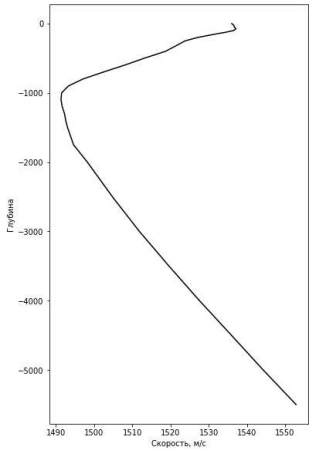

Рис.1.1 Профиль скорости звука ВРСЗ

В данной статье был предложен подход, основанный на алгоритмах машинного обучения. Похожая работа была проделана в [1]. Данная работа была дополнена методом, основанном на применении нейронных сетей.

В работе используются методы unsupervised learning (обучение без учителя). При таком подходе известно только описание объектов (обучающей выборки), и требуется обнаружить внутренние закономерности, зависимости между объектами. Такой подход позволяет снизить размерность за счет выделения только наиболее важной информации из данных.

Для решения данной задачи были рассмотрены следующие методы:

1. Метод главных компонент.
2. Метод k-means (к средних).
3. K-SVD (Singular value decomposition).
4. Нейронные сети.

2. ПОСТРОЕНИЕ МОДЕЛИ И ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ

Все модели, представленные далее, были обучены на выборке из 112 профилей ВРСЗ. Все результаты были получены на тестовой выборке из 38 профилей, размерность исходного профиля - 33. Оценивается среднеквадратическая ошибка (mean square error, MSE). Результаты были получены средствами языка python и библиотек для машинного обучения - scikit learn, для обучения нейронных сетей - Tensor Flow (keras api), и специальной библиотекой для метода k-svd - ksvd.

Анализ был произведен для небольшой окрестности (широта от 10 до 32, долгота от -70 до -40) Индийского океана по данным за январь.

Метод главных компонент (principal component analysis, PCA)

Данный метод позволяет кодировать точки в пространстве меньшей размерности. Для каждой точки $x^i \in \mathbb{R}^n$, $i = 1..N$, где N это количество элементов в выборке, требуется найти соответствующий ей кодированный вектор $c^i \in \mathbb{R}^k$. Если k меньше n, то для хранения коди-

рованных точек потребуется меньше памяти, чем для исходных. Требуется найти функцию кодирования $f(x) = c$ и функцию декодирования $x \approx g(f(x))$.

Используя метод главных компонент, задача сводится к тому, чтобы найти такое ортогональное преобразование в новую систему координат, в которой выборочная дисперсия данных вдоль первой координаты максимальна. Выборочная дисперсия вдоль второй координаты максимальна при условии ортогональности первой координате, выборочная дисперсия вдоль n-ой координаты максимальна при условии ортогональности всем остальным координатам.

Возьмем такой набор векторов X, в котором каждый вектор имеет среднее 0. Если это не так, центрирования легко добиться, вычтя среднее из всех примеров на этапе предварительной обработки.

Несмещенная выборочная ковариационная матрица, ассоциированная с X, определяется по формуле :

$$Var(x) = \frac{1}{N-1} X^T X \qquad (2.1)$$

PCA находит представление (посредством линейного преобразования) $z = W^T x$, для которого $Var(z)$ диагональная матрица. Отсюда легко показать, что главные компоненты матрицы X определяются собственными векторами $X^T X$. Таким образом $X^T X = W \Lambda W^T$ [4].

Для сжатия профиля ВРСЗ до размерности k, достаточно взять первые k главных компонент и разложить по ним данный профиль.

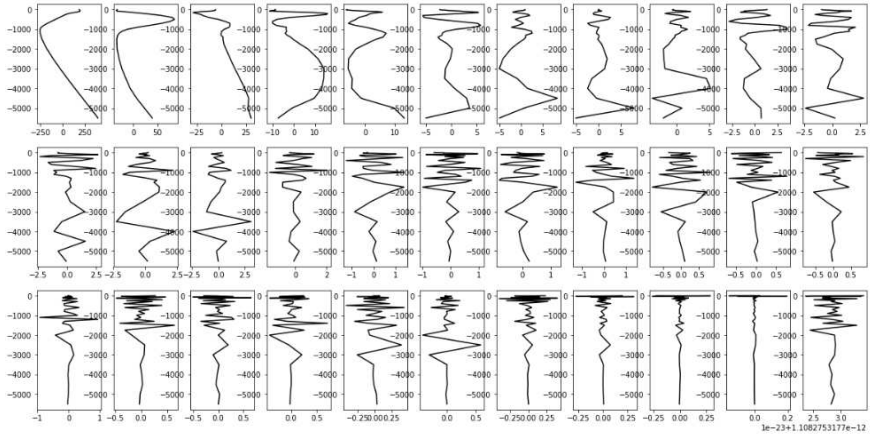

Рис.2.1 Главные компоненты профилей скорости звука

На рисунке 2.1 представлен набор главных компонент для профилей скорости звука.

На рисунке 2.2 приводится сравнение исходного вектора из тестовых данных и его аппроксимации по 1, 3 и 5 главным компонентам. В таблице 2.1 представлено среднеквадратическое отклонение исходного тестового вектора от аппроксимированного.

Таблица 2.1

Количество компонент	Среднеквадратическое отклонение
1	2,03
3	1,486
5	0,15

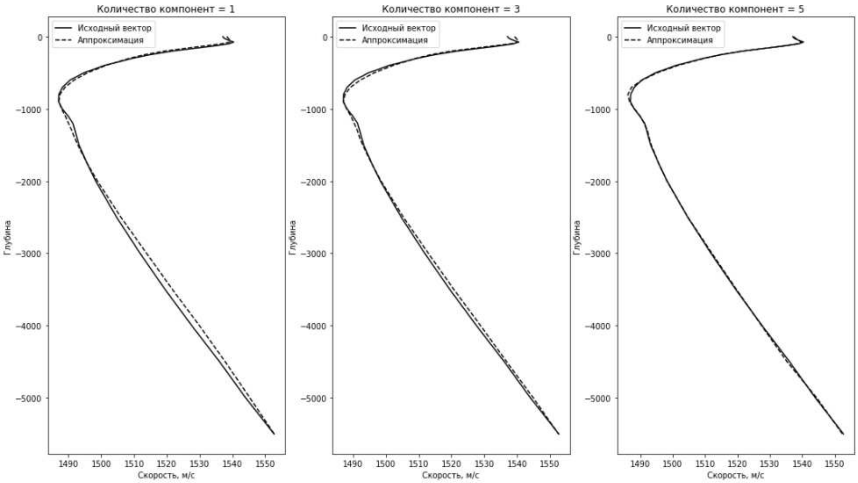

Рис.2.2. Исходный вектор ВРСЗ и его аппроксимация методом PCA

K-means

В машинном обучении метод k-means является методом кластеризации данных (присвоение каждому элементу выборки метку из конечного числа кластеров заранее неизвестных). Несмотря на это, можно провести параллели между алгоритмом k-means и PCA. Если метод PCA пытается представить данные в виде суммы главных компонент, то метод k-means напротив, пытается представить каждую точку данных в пространстве, используя центр кластера. Разбивая n-мерное пространство выборки ВРСЗ на k кластеров, каждый вектор ВРСЗ кодируется в k-мерном пространстве по принципу one-hot-encoding, то есть кодируется вектором из k элементов, в котором на i-ом месте стоит единица, а все остальные нули, где i – номер кластера к которому принадлежит данный профиль ВРСЗ.

В таблице 2.2 представлено среднеквадратическое отклонение исходного вектора из тестовых данных от вектора, являющегося центром ближайшего кластера.

Таблица 2.2

Количество кластеров	Среднеквадратическое отклонение
1	35,22
3	5,64
5	7,38
8	0,83
10	3,38

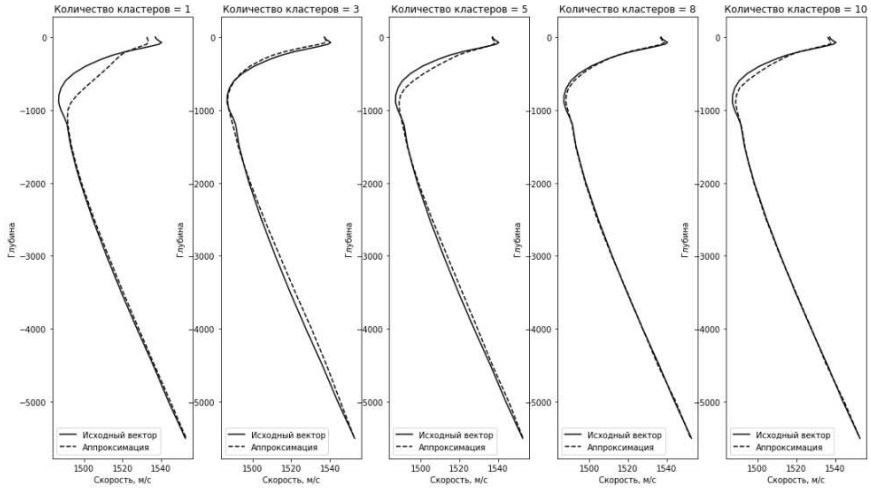

Рис.2.3. Исходный вектор ВРСЗ и его аппроксимация методом k-means

Данный метод кодирования хоть и не самый точный, но зато позволяет получить очень разреженное представления данных, что во многих задачах бывает очень практично.

K-SVD

K-SVD – еще один способ получить разреженное представления данных. Это частный случай метода под названием dictionary learning. Он позволяет эффективно находить словарь (набор базисных векторов ВРСЗ из обучающей выборки) с помощью SVD разложения и является своего рода обобщением метода k-means. В отличии от k-means, K-SVD позволяет получать кодированный вектор с заранее заданным количеством ненулевых элементов. В методе K-SVD решается следующая задача оптимизации:

$$\min_{X,D}\{\|Y - DX\|_F^2\}, \text{ при условии } \|x_i\|_0 \le T_0, \text{ для } \forall i \in 1..N, \qquad (2.2)$$

где

$Y \in \mathbb{R}^{n\times N}$ - обучающий набор профилей ВРСЗ, $D \in \mathbb{R}^{n\times k}$ - матрица словарь - составленная из k обучающих профилей ВРСЗ, $X \in \mathbb{R}^{k\times N}$ - кодированные профили ВРСЗ, F - норма Фробениуса (корень из суммы квадратов всех элементов матрицы), $\|.\|_0$ - l_0 норма вектора (количество ненулевых элементов в векторе), T_0 - максимальное количество ненулевых элементов в кодированном векторе ВРСЗ.

Данная задача решается итеративным поиском при помощи SVD разложения, более подробный алгоритм изложен в [2]. Мы для решения воспользуемся готовой библиотекой языка python – ksvd.

K-SVD является обобщением метода k-means, так как при $T_0 = 1$, K-SVD эквивалентен методу k-means. Восстановленный вектор по методу K-SVD уже является не центром одного кластера, как это было в k-means, а линейной комбинацией нескольких таких кластеров.

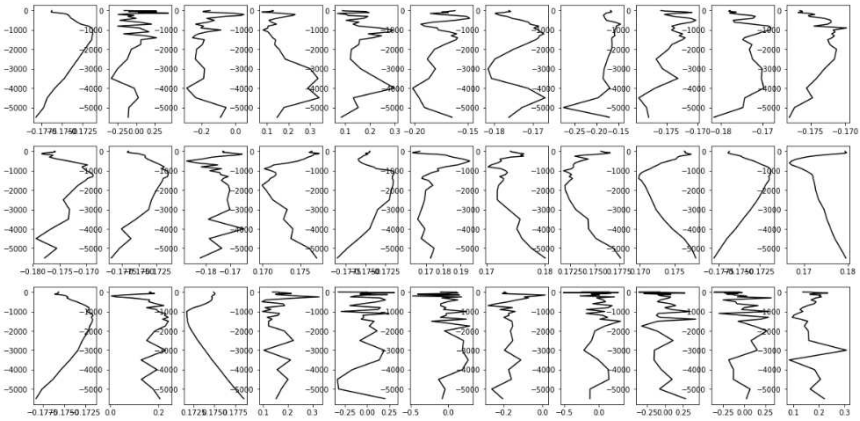

Рис.2.4. Словарь, полученный методом K-SVD

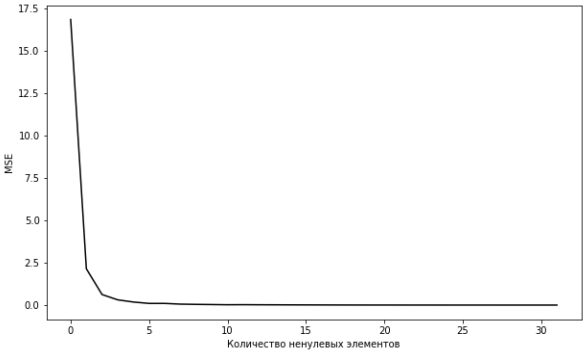

Рис.2.5. Зависимость ошибки аппроксимации от количества ненулевых элементов.

Как видно из графика кривой ошибки обучения, показанного на рисунке 2.5, точность перестает заметно увеличиваться уже при $T_0 = 5$.

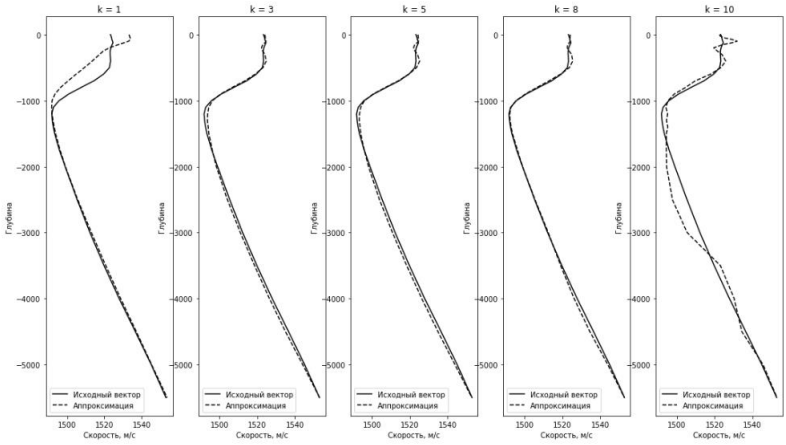

Рис.2.6. Исходный вектор ВРСЗ и его аппроксимация методом K-SVD

Таблица 2.3

K	Среднеквадратическое отклонение
1	49.7
3	1
5	1.07
8	0,82
10	10.72

Нейронные сети

Как уже говорилось ранее, в модели PCA, для сжатия размерности данных требуется найти такую функцию кодирования $f(x) = c$, $x \in \mathbb{R}^n$, $c \in \mathbb{R}^l$ ($l < n$) и функцию декодирования $x \approx g(f(x))$.

Теорема Цыбенко

Искусственная нейронная сеть прямой связи с одним скрытым слоем может аппроксимировать любую непрерывную функцию многих переменных с любой точностью. Условиями являются: достаточное количество нейронов скрытого слоя [3].

Данная теорема утверждает, что мы можем аппроксимировать функцию $f(x)$ и $g(c)$ с любой точностью.

Для данного типа задач существует специальная архитектура нейронной сети под названием autoencoder (автокодировщик). Она состоит из двух частей: кодирование и декодирование.

На рисунке 2.7 приведена архитектура автокодировщика с одним скрытым слоем кодирования и одним скрытым слоям декодирования.

Методом обратного распространения ошибки, нейронную сеть можно обучить кодировать данные в пространство меньшей размерности (нужно всего лишь минимизировать среднеквадратическое отклонение между исходными векторами и векторами на выходе нейронной сети). Выбирая количество нейронов на скрытом слое С, будем задавать размерность пространства, в которое перейду профили ВРСЗ.

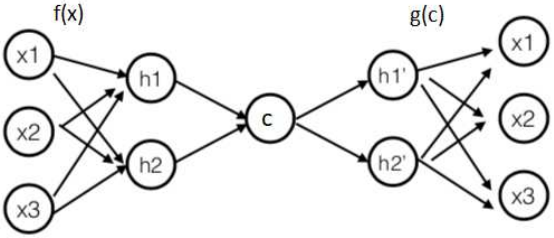

Рис.2.7 Архитектура autoencoder (автокодировщик)

После обучения нейронной сети, обрезав вторую часть, которая отвечает за декодирование данных, и оставив только часть, отвечающую за функцию кодирования $f(x)$, будем подавать на вход сети профили ВРСЗ и на выходе будем получать закодированные вектора размерности скрытого слоя С. Так как у нас остался обученный декодер $g(c)$, мы в любой момент можем восстановить (с потерей точности конечно) профили ВРСЗ прогнав их через сеть декодера.

Таблица 2.4

Размер скрытого слоя С	Среднеквадратическое отклонение
1	1.18
3	2.12
5	1.6
10	1.3

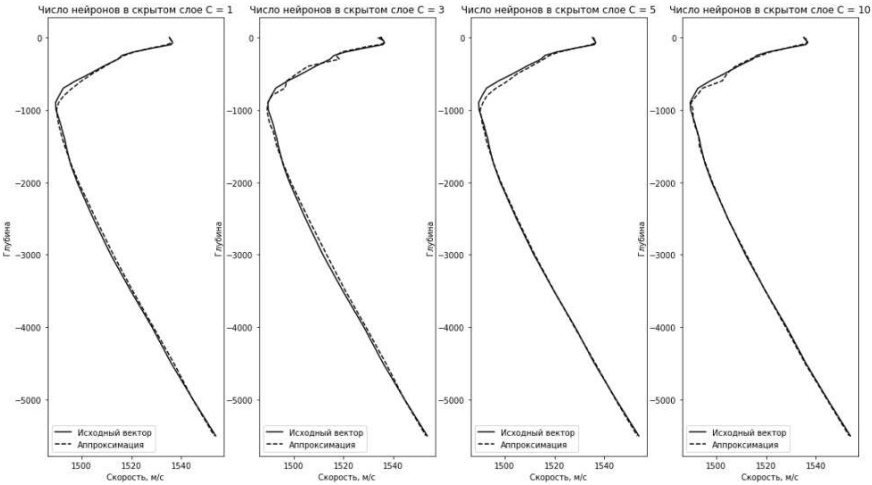

Рис2.8. Исходный вектор ВРСЗ и его аппроксимация с использованием autoencoder.

Сеть обучалась методом обратного распространения ошибки на 150 эпохах и алгоритмом оптимизации adam.

3. ЗАКЛЮЧЕНИЕ

Сравним на одном графике результаты всех четырех моделей:

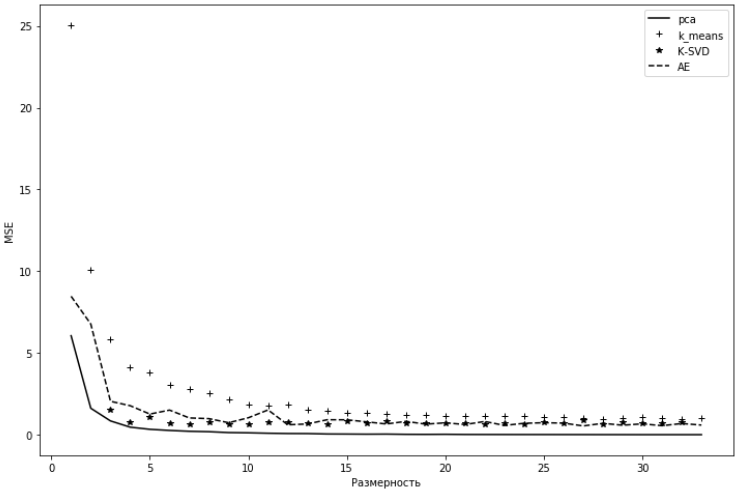

Рис3.1. Сравнение четырех моделей

Как видно из графика, метод PCA лучше всех проявил себя при решении данной задачи. Как и ожидалось, K-SVD более точнее, чем k-means, так как является его обобщением. Хоть PCA и показал более точные результаты на тестовых данных, все равно есть смысл применять метод K-SVD, за счет разреженных кодируемых данных.

БЛАГОДАРНОСТИ
Автор благодарит за помощь в исследовании научного руководителя проекта М.В. Лебедева.

ЛИТЕРАТУРА

1. Dictionary learning of sound speed profilesю. Michael Biancoa, Peter Gerstoft.
2. K-SVD and its Non-Negative Variant for Dictionary Design. Michal Aharon Michael Elad Alfred M. Bruckstein. Department of Computer Science Technion—Israel Institute of Technology Technion City, Haifa 32000, Israel
3. Approximation by Superpositions of a Sigmoidal function, Mathematics of Control Signals and Systems. Cybenko, G. V.
4. Deep learning. Ian Goodfellow, Yoshua Bengio, Aaron Courville.
5. Introduction to Machine Learning with Python. Andreas Müller, Sarah Guido.
6. http://oalib.hlsresearch.com/Modes/AcousticsToolbox/manual_html/node8.html

Работа поступила 20.02.2019г.