Çukurova Üniversitesi
Mühendislik Fakültesi Dergisi
Çukurova University
Journal of the Faculty of Engineering
ISSN: 2757-9255 CİLT/VOLUME: 39 SAYI/ISSUE: 4 ARALIK/DECEMBER 2024

Comparison of Feature Extraction Methods in High Dimensional Time Series

Emre KILINÇ [1,a]

[1]*Ağrı İbrahim Çeçen University, Patnos Vocational College, Department of Computer Technologies, Ağrı, Türkiye*

[a]***ORCID**: 0000-0002-5250-9322*

Article Info
Received : 19.08.2024
Accepted : 23.12.2024
DOI: 10.21605/cukurovaumfd.1606090

Corresponding Author
Emre KILINÇ
ekilinc@agri.edu.tr

Keywords
Machine learning
Feature extraction
Time series
Dimension reduction

***How to cite:** KILINÇ, E., (2024). Comparison of Feature Extraction Methods in High Dimensional Time Series. Cukurova University, Journal of the Faculty of Engineering, 39(4), 991-997.*

ABSTRACT

Working with high-dimensional datasets increases the workload on machine learning models. Therefore, before making predictions, the most meaningful data points in the entire data set must be determined. It is highly important to improve model performance, especially in the field of machine learning. For this reason, five feature selection methods—Mutual Information, Principal Component Analysis, Chi-square, Information Gain, and Variance Thresholding—commonly used in the literature, were tested on the 14400 feature data set obtained with a system previously proposed to determine the sand, silt and clay ratios in the soil. The success of these five methods is presented comparatively using R-square (R^2) and Mean Absolute Error (MAE) metrics. The best results were obtained with the Information Gain method for sand ($R^2 = 0.44$), with Chi-square for silt ($R^2 = 0.17$), and with Variance Thresholding for clay ($R^2 = 0.61$).

Yüksek Boyutlu Zaman Serilerinde Özellik Çıkarma Yöntemlerinin Karşılaştırılması

Makale Bilgileri
Geliş : 19.08.2024
Kabul : 23.12.2024
DOI: 10.21605/cukurovaumfd.1606090

Sorumlu Yazar
Emre KILINÇ
ekilinc@agri.edu.tr

Anahtar Kelimeler
Makine öğrenmesi
Özellik çıkarma
Zaman serisi
Boyut indirgeme

***Atıf şekli:** KILINÇ, E., (2024). Comparison of Feature Extraction Methods in High Dimensional Time Series. Cukurova University, Journal of the Faculty of Engineering, 39(4), 991-997.*

ÖZ

Yüksek boyutlu veri setlerinde, makine öğrenmesi ile çalışmak iş yükünde artışa sebep olmaktadır. Bu nedenle tahminleme işlemleri yapılmadan önce, tüm veri seti içerisindeki en anlamlı veri noktalarının belirlenmesi gerekmektedir. Özellikle makine öğrenmesi alanında model performansını artırmak için kritik öneme sahiptir. Bu nedenle daha önce topraktaki kum, silt ve kil oranlarını belirlemek amacıyla önerilen bir sistemle elde edilen 14400 özellikli veri seti üzerinde, literatürde sıklıkla kullanılan Karşılıklı Bilgi, Temel Bileşen Analizi, Ki-kare, Bilgi Kazancı ve Varyans Eşiği Belirleme özellik seçme metotları denenmiştir. Bu 5 metodun başarı sonuçları R-kare (R^2) ve Ortalama Mutlak Hata (OMH) cinsinden karşılaştırmalı olarak sunulmuştur. En iyi sonuçlar kum için Bilgi Kazancı metodu ile ($R^2 = 0.44$), silt için Ki-kare ile ($R^2 = 0.17$), kil için Varyans Eşiği Belirleme ile ($R^2 = 0.61$) elde edilmiştir.

1. INTRODUCTION

Feature extraction is the process of converting raw data into a more informative format so that machine learning algorithms can be processed effectively and is a critical process in machine learning [1-3]. These extracted features are essential for functions such as classification, pattern recognition and understanding complex processes [4]. In fields such as monitoring, image recognition, and structural engineering, selecting and extracting relevant features is important for accurate analysis and decision making [5,6]. Advanced feature extraction methods show promise in automating the feature extraction process and improving classification accuracy [7-9].

Feature extraction is also a fundamental process in the analysis of time series signals in various fields. Identifying meaningful features by reducing the size of time series data increases the performance and accuracy of machine learning models put forward [10,11]. Various feature extraction methods have been developed in the literature to extract relevant information from datasets, such as Mutual Information (MI), Principal Component Analysis (PCA), Chi-square, Information Gain (IG) and Variance Threshold (VT) methods [12,13]. MI and IG measure the dependence between variables and uncertainty between features and are used in text processing, biomedical, etc [12]. PCA is a technique used to reduce the dimension of a dataset by mapping data to a new coordinate system to capture the most significant variance [14]. Chi-square is used in tasks such as sentiment analysis and disease prediction by assessing the relationship between features and target variables [15]. Variance Thresholding is ideal for eliminating low-variance features that provide insignificant information and to decrease the data workload. After all, researchers can improve the efficiency and accuracy of various applications by using advanced feature extraction techniques tailored to the characteristics of time series data.

In this study, experiments were carried out on a data set consisting of 66 observations and 14400 features using the 5 feature extraction algorithms mentioned above, and the results are presented comparatively in detail in the following sections.

2. MATERIAL AND METHODS

2.1. Dataset

In the study, time series signals obtained with the system called USTA, which was introduced in a previous study, were used [16]. In the previous study, the amplitude of sound transmitted through the soil-water mixture was measured with a pair of transmitter-receiver sensor. For 80 soil samples, a time series dataset of 14400 features, ranging from 0 to 5 volts, was created at a frequency of 2Hz. These data were processed with various machine learning methods and sand, silt and clay predictions were made. In this study, soils with 3 or more samples of the same class in the dataset were selected to avoid uninterpretable results during the training phase. In total, dataset consists of 66 observations and 14400 columns were used.

2.2. Mutual Information (MI)

MI is a statistical measurement method used to measure the amount of information shared between two variables. The basic idea is to evaluate how much knowing the value of a particular attribute reduces uncertainty about the target variable. Entropy for a single variable is calculated using the probability distribution of the values of that variable and its formula is as in Equation 1.

$$H(X) = -\sum p(x) log_2^{p(x)} \tag{1}$$

Here $p(x)$ refers to the marginal probability distribution. Then MI for each variable is calculated using Equation 2.

$$MI(X;Y) = H(X) + H(Y) - H(X,Y) \tag{2}$$

MI refers to the shared information between two variables, and $H(X,Y)$ refers to the calculated common entropy value. Features with higher MI scores are considered to have a stronger relationship with the target variable.

2.3. Principal Component Analysis (PCA)

PCA aims to transform the original features into a new set of uncorrelated features that capture the maximum variance in the data. The steps are as follows;

- Data is centered by subtracting the mean of each feature
- Covariance matrix of the centered data is calculated.
- Eigenvalue decomposition of the covariance matrix is performed
- The principal components are ranked according to their corresponding eigenvalues.
- Original data is projected onto the selected principal components.

The biggest advantage of PCA is its ability to reduce dimensionality while minimizing information loss. A big disadvantage of PCA is that, being a linear method, it may perform poorly with datasets that have complex, non-linear relationships. However, its applicability extends to non-linear problems to capture non-linear relationships in data by mapping it into a higher-dimensional feature space where linear separability can be achieved [17].

2.4. Chi-square Test

Chi-square test is used to determine whether there is a significant relationship between two categorical variables. One being the feature, the other is the target. It is calculated as in Equation 3.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \tag{3}$$

Here O_i represents the observed frequency for a category, and E_i represents the expected frequency assuming independence. Features with higher Chi-square scores have a stronger relationship with the target class.

2.5. Information Gain (IG)

The origin of IG is based on the concept of entropy. In this aspect, it is similar to the MI method. In the context of feature extraction, entropy represents the uncertainty of the target variable. The IG of a feature indicates how much knowing that feature reduces the entropy of the target variable. Steps are as follows;

- Entropy of the target variable is calculated to measure the overall uncertainty in the dataset.
- For each feature, conditional entropy of target is calculated.
- To obtain the IG for each feature, conditional entropy is subtracted from the original entropy.
- Features are ranked based on their IG scores.
- The top-ranking features are used.

It is an effective feature extraction method that can be used quickly and practically. The disadvantage is that if the data set consists of continuous features, the data set must be discretized.

2.6. Variance Thresholding (VT)

Variance thresholding works with the assumption that features with low variance across samples do not carry much information about the target variable. The variance of the dataset with m samples and n features is found mathematically as in Equation 4.

$$Var(X_j) = \frac{1}{m}\sum_{i=1}^{m}(x_{ij} - \mu_j)^2 \tag{4}$$

Here, j represents the feature index and x_{ij} represents the value of the j^{th} feature for the i^{th} sample. μ_j is the mean of the j^{th} feature across all samples. In variance thresholding, a threshold value (t) is set, and only features with variance greater than t are considered.

2.7. Evaluation Metrics

In this study, R-square (R^2) and Mean Absolute Error (MAE) performance metrics were used to see the effect of 5 different feature extraction methods on prediction success.

The R^2 metric is used to determine the proportion of variance in the dependent variables. In mathematics, its expression is as in Equation 5.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \quad (5)$$

Here y_i represents the real values, $\bar{y}$ is the average of the real values, and n is the number of samples. R^2 value is between 0 and 1. A higher value indicates that the model is more successful.

MAE value is used to measure the absolute difference between actual values and predicted values. Its mathematical expression is as in Equation 6.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (6)$$

Here, y_i represents the actual values, $\hat{y}_i$ represents the predicted value, and n indicates the number of samples. A lower MAE value means a better model.

3. EXPERIMENTS AND RESULTS

After the time series data of 66 soil samples in the data set were subjected to the MI feature extraction method, the target variables and features were divided into bins for entropy calculation. MI was calculated separately for each feature based on sand, silt and clay ratios. MI scores were then averaged to obtain a joint MI score for each feature. For regression analysis, the best 50 features were selected according to the combined MI scores, and 65 soil sample data were given to the Random Forest (RF) method with the leave-one-out cross validation method. The prediction results made with the RF model with 100 random trees, were recorded, and the R^2 values and prediction successes are given in Figure 1.

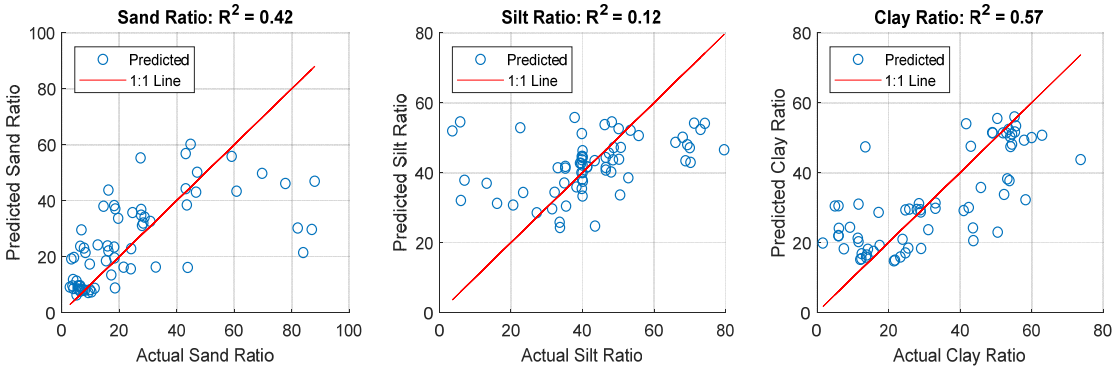

Figure 1. Sand, silt and clay estimation results using the MI feature extraction method on 66 soil samples

As can be seen in Figure 1, R^2 values were found to be 0.42 for sand, 0.12 for silt and 0.57 for clay. Success rates in terms of R^2 and MAE are given in the table below.

Table 1. R^2 and MAE values of the prediction results using the features obtained with the MI method

	Sand	**Silt**	**Clay**
R^2	0.42	0.12	0.57
MAE	11.57	10.74	9.2

R^2 and MAE values are given comparatively in Table 1. In addition to R^2 values, MAE values are around 10% and are within acceptable limits, just like in the traditional hydrometer method.

3.2. Experiments with Principal Component Analysis (PCA)

PCA was performed on the normalized time series data and principal components explaining 95% of the variance were selected. Then, predictions were made using the leave-one-out cross-validation method with these selected principal components. The comparative estimation results are provided in Table 2.

Table 2. R^2 and MAE values of the prediction results using the features obtained with the PCA method

	Sand	Silt	Clay
R^2	0.14	0.10	0.13
MAE	12.02	14.06	12.52

As seen in the Table 2, R^2 values for sand, silt and clay were found to be 0.14, 0.10 and 0.13, respectively. On the other hand, MAE values are well above the 10% margin of error. It appears that PCA produces worse results than MI when using such a data set.

3.3. Experiments with Chi-square

Both the features and the target variables (sand, silt, clay ratios) were decomposed into bins to apply the Chi-square test, which works with categorical data. For each feature, contingency tables were created to examine the relationship between the discretized feature and the target variable. The chi-square statistic is calculated separately for each feature based on sand, silt, and clay ratios. A combined Chi-square score is obtained by averaging the Chi-square scores for each feature. Then, the best 50 features were selected according to these combined scores, and the selected features were given as input to the RF with the leave-one-out cross validation method. The prediction rates obtained are given in Table 3.

Table 3. R^2 and MAE values of the prediction results using the features obtained with the Chi-square method

	Sand	Silt	Clay
R^2	0.30	0.17	0.51
MAE	13.63	10.22	10.24

When Table 3 is examined, although the R^2 values seem to be low, when we look at the MAE values, it can be seen that mediocre estimates are obtained for silt and clay. Although not as good as the results obtained with MI, it can be seen that better results are obtained than PCA analysis.

3.4. Experiments with Information Gain (IG)

First, the target variables (sand, silt, clay ratios) and time series features are divided into bins to implement the IG calculation. Then, by calculating the entropy value for each target variable and each discretized feature, the common entropy between each target and feature was determined. The IG for each feature was calculated by subtracting the joint entropy from the target's entropy. Finally, the determined features were given to the RF method with the leave-one-out method and predictions were made. Table 4 shows these estimation results.

Table 4. R^2 and MAE values of the prediction results using the features obtained with the IG method.

	Sand	Silt	Clay
R^2	0.44	0.13	0.57
MAE	10.18	10.68	9.11

Table 4 shows the prediction success achieved when the IG feature extraction method was used. The IG feature extraction method yielded the best results for predicting sand, silt, and clay. Especially when we look at the MAE values, it can be seen that all predictions are within acceptable limits.

3.5. Experiments with Variance Thresholding (VT)

As a first step, the variance of each feature in the normalized time series data was calculated. Features that were above a predefined variance threshold (threshold = 0.01 in this case) were selected and other features were eliminated. These selected features were given as input to the RF method and predictions were made.

The estimation results obtained with the Variance Thresholding feature extraction method are given in Table 5.

Table 5. R^2 and MAE values of the prediction results using the features obtained with the VT method.

	Sand	**Silt**	**Clay**
R^2	0.41	0.11	0.61
MAE	10.1	12.06	8.64

When Table 5 is examined, the MAE values obtained are at an acceptable level except for silt estimation. It produced the best results compared to other methods, in terms of clay estimation. The success rates obtained by all methods are presented comparatively in Table 6.

Table 6. R^2 and MAE values of prediction results using features obtained by MI, PCA, Chi-square, IG and VT methods.

		Sand	**Silt**	**Clay**	**Average**
MI	R^2	0.42	0.12	0.57	0.37
	MAE	11.57	10.74	9.2	10.5
PCA	R^2	0.14	0.10	0.13	0.12
	MAE	12.02	14.06	12.52	12.86
Chi-square	R^2	0.30	**0.17**	0.51	0.33
	MAE	13.63	**10.22**	10.24	11.37
IG	R^2	**0.44**	0.13	0.57	**0.38**
	MAE	**10.18**	10.68	9.11	**9.9**
VT	R^2	0.41	0.11	**0.61**	0.37
	MAE	10.1	12.06	**8.64**	10.2

Table 6 shows the success rates of sand, silt and clay predictions obtained with 5 different feature extraction methods. The highest success rates were obtained with IG for sand (R^2 = 0.44, MAE = 10.18), with Chi-square for silt (R^2 =0.17, MAE = 10.22) and with VT for clay (R^2 = 0.61, MAE = 8.64). The best results can be achieved by using these three methods in combination. However, if a single method that gives the optimum result is to be chosen, the best option seems to be the IG method. As noted in the "Average" column of Table 6, IG already produces the best results for sand estimation, and comes close to the best results for silt and clay estimations with very small compromises. Using a single feature extraction method will also reduce the workload and complexity of the prediction system to be developed.

4. CONCLUSIONS

In large-scale data (especially time series), it is very important to select the most meaningful signals instead of all signals. In this way, the dataset can be used more effectively in machine learning methods by reducing its size. The aim here is to work faster with less data, compromising the prediction success as little as possible. In this study, 5 feature extraction methods frequently used in the literature were applied on a 14400-column (feature) time series dataset of 66 soil samples obtained with the previously introduced USTA device. A key strength of this research lies in its focus on evaluating feature extraction methods within the context of high-dimensional soil data rather than aiming solely for predictive accuracy. By employing the Random Forest algorithm as a benchmarking tool, the study provided a controlled environment to compare the effectiveness of each method. This approach ensures that the reported results are robust and generalizable, offering practical guidance for future research and applications. First, future work could explore integrating feature extraction methods with ensemble machine learning models to enhance accuracy without notably increasing computational costs. Second, future research could investigate the applicability of these methods to other types of soil data, potentially incorporating spatial or environmental factors to broaden the scope of their utility. Third, leveraging advanced optimization techniques, such as evolutionary algorithms or neural architecture search, may further enhance feature selection, tailoring the process to specific datasets or applications.

5. REFERENCES

1. Wangni, J., Chen, N., 2016. Nonlinear feature extraction with max-margin data shifting. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1), 10299.
2. Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.
3. Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer Series in Statistics. Springer, New York, NY, 745.
4. Ren, S., Zhang, X., Li, H., Chu, G., Chen, D., Bai, H., Hu, C., 2022. Interpretable feature extraction for the numerical particle system. In B.H.V. Topping, & P. Iványi (Eds.), Proceedings of the Eleventh International Conference on Engineering Computational Technology. Civil-Comp Press, Edinburgh, UK.
5. Alegeh, N., Thottoli, M., Mian, N.S., Longstaff, A.P., Fletcher, S., 2021. Feature extraction of time-series data using DWT and FFT for ballscrew condition monitoring. Advances in Transdisciplinary Engineering.
6. Wang, Y., 2022. Malicious URL detection: An evaluation of feature extraction and machine learning algorithm. Highlights in Science, Engineering and Technology, 23, 117-123.
7. Qian, X., Zhang, H., Yang, C., Wu, Y., He, Z., Wu, Q.-E., Zhang, H., 2018. Micro-cracks detection of multicrystalline solar cell surface based on self-learning features and low-rank matrix recovery. Sensor Review, 38(3), 360-368.
8. Xu, Y., Yin, K., Zhang, J., Yao, L., 2008. A spatiotemporal approach to N170 detection with application to brain-computer interfaces. 2008 IEEE International Conference on Systems, Man, and Cybernetics.
9. Doraikannan, S., Selvaraj, P., Burugari, V.K., 2019. Principal component analysis for dimensionality reduction for animal classification based on LR. International Journal of Innovative Technology and Exploring Engineering, 8(10), 1118-1123.
10. Lin, J., Li, H., Zhou, C., Li, W., Shao, X., 2023. Autoencoder-based feature extraction for power time series data considering social information. Eighth International Conference on Electromechanical Control Technology and Transportation (ICECTT 2023).
11. Liu, H., Motoda, H., 1998. Feature selection for knowledge discovery and data mining. The Springer International Series in Engineering and Computer Science. Springer, US.
12. Liu, S., Tang, B., Chen, Q., Wang, X., Fan, X., 2015. Feature engineering for drug name recognition in biomedical texts: Feature conjunction and feature selection. Computational and Mathematical Methods in Medicine, 2015, 1-9.
13. Bishop, C.M., 2006. Pattern recognition and machine learning. Information Science and Statistics. Springer, New York, NY, 778.
14. Venkatesh, R., Anantharajan, S., Gunasekaran, S., 2023. Multi-gradient boosted adaptive SVM-based prediction of heart disease. International Journal of Computers Communications & Control, 18(5), 4994.
15. Yusliani, N., Aruda, S.A.Q., Marieska, M.D., Saputra, D.M., Abdiansah, A., 2022. The effect of chi-square feature selection on question classification using multinomial naïve Bayes. Sinkron, 7(4), 2430-2436.
16. Orhan, U., Kilinc, E., Albayrak, F., Aydin, A., Torun, A., 2022. Ultrasound penetration-based digital soil texture analyzer. Arabian Journal for Science and Engineering, 47(8), 10751-10767.
17. Schölkopf, B., Smola, A.J., Müller, K., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation, 10(5), 1299-1319.