Journal of
Computer & Electrical and Electronics Engineering Sciences

Original Article

DOI: 10.51271/JCEEES-0010

Machine learning methods on quantized vectors

Ufuk Tanyeri[1] Taner Dindar[2] Yunus Kökver[3] Nuri Furkan Koçak[2]

[1]Department of Computer Technologies, Nallıhan Vocational School, Ankara University, Ankara, Turkey
[2]Department of Electronic Technologies, Nallıhan Vocational School, Ankara University, Ankara, Turkey
[3]Department of Computer Technologies, Elmadağ Vocational School, Ankara University, Ankara, Turkey

Cite this article: Tanyeri U, Dindar T, Kökver Y, Koçak NF. Machine learning methods on quantized vectors. *J Comp Electr Electron Eng Sci.* 2023;1(2):46-49.

Corresponding Author: Ufuk Tanyeri, ufuktanyeri@ankara.edu.tr

Received: 15/09/2023 ◆ **Accepted**: 19/10/2023 ◆ **Published**: 31/10/2023

ABSTRACT

Vector quantization is one of the important issues in digital images. There are many studies conducted on quantized vectors or images. On the other hand, machine-learning approaches are a popular issue today. In this study, the classification performances of machine learning approaches on reduced image vectors are examined. Firstly, Corel 1K data set were reduced to 64 colors with octree and histogram feature vectors extracted. Classification was carried out using various machine learning approaches on the relevant vectors. As a result of the classification, the success of the methods was examined.

Keywords: Vector quantization, machine learning, corel 1k, classification

INTRODUCTION

Vector quantization (VN) is a technique used in signal processing and data compression approaches. While analysis, modification, or improvement of analog or digital signals is carried out with signal processing, data size is reduced while preserving important features with data compression. Lloyd (1957) laid the scalar-based foundation of VN in the field of signal processing (Lloyd, 1957), while Forgy used it in vector form (Forgy, 1965). In fact, the issue of optimal quantization of a vector space is called Dirichlet tessellation in two- and three-dimensional spaces and Voronoi tessellation in arbitrary-dimensional spaces (Lejeune Dirichlet, 1850; Voronoi, 1908). VN is a standard in digital signal processing today. With this method, a large set of data points is represented by a smaller set of reference points known as centroids or codebook vectors. Each data point is then assigned to the nearest center. This way, the data is effectively quantified. If the input data is finite, the bulk calculation method called Linde-Buzo-Gray (LBG) is used (Linde et al, 1980). There is also a k-means classification that is widely used in the literature (Gersho, 1982; Gray, 1984; Makhoul et al, 1985). Octree is another method used to divide a three-dimensional space into smaller and more manageable parts (Meagher, 1982). It is particularly used in computer graphics, 3D modeling, geographic information systems (GIS), game development, and similar fields.

Today, the increase in the number of digital images brings with it the problem of storing them. One solution is to quantize the pixel values in the image into a smaller set of values using VN. VN can greatly reduce the storage requirements of images while preserving essential features. Additionally, the size of voice and many other types of data can also be reduced. Pattern recognition is another area where VN is used. Pattern recognition is a versatile method that helps understand, classify, group data, and predict future events (Li et al, 2021; Serey et al, 2023; Mantel, 1974). In this field, since the performance of VN depends on the design of the code book, the selected distance measure, and the specific application area, it is preferred as an alternative to other methods such as neural networks.

Another important issue, such as quantization of data, is classification. Classification is based on machine learning algorithms and is used to separate input data points into predefined classes or categories. These techniques leverage input training data to predict the likelihood that subsequent data will fall into one of the predetermined categories. This way, it can find the same pattern (similar words or emotions, number sequences, etc.) in future datasets.

The field of classification studies is very extensive, and depending on the dataset you are working with, you can employ a variety of techniques. Logistic regression, naive bayes, random forest, gradient boosting, k-nearest neighbors (KNN), decision trees and support vector machines (SVM) are common approaches. Another alternative classification method logistic regression (LR) is used for binary classification. Here, the probability of a certain entry point belonging to a certain class is modeled using the logistic function (Kononenko, 1989). Naive bayes (NB) is based on the bayes theorem and assumes that the presence of a particular feature in a class is not correlated with the presence of other features (Palmer et al, 1979). The random forest (RF) model uses decision trees. RF builds each tree on a bootstrapped subset of the data and combines the predictions to improve accuracy and control overfitting (Friedman, 2001). Gradient boosting (GB) is a frequently used ensemble learning method in the field of machine learning. The related method aims to create a stronger learner by combining weak learners (usually decision trees) (Fukunaga et al, 1975). Variations of the GB approach include XGBoost (extreme gradient boosting), LightGBM (light gradient there are libraries such as boosting machine) and CatBoost. K-nearest neighbors (KNN) is a machine learning method used as a classification or regression algorithm. The basic idea is to use most data points around a data point to classify a new data point or make a prediction. K-NN is among the supervised learning algorithms and is used in data mining, pattern recognition and classification problems (Magee, 1964). Decision trees (KA) are another classification and regression algorithm used in data mining and machine learning fields. The main purpose is to classify new data samples or make predictions by analyzing the dataset. Decision trees use a tree structure containing a set of decision nodes and leaf nodes (Vapnik et al, 1996). Finally,s upport vector machines (SVM), one of the classification algorithms, aims to find a hyperplane that best separates different classes in the input feature field and maximizes the margin between them.

In this study, feature vectors were first extracted from the corel 1K data set by reducing it to 64 colors using the octree method. Then, the relevant vectors were classified with machine learning-based approaches. The motivation of the study is to examine the classification success of machine learning approaches of images reduced to 64 colors with the octree reduction technique.

This paper is structured as follows. Information on vector quantization is provided in section 2. Following that, section 3 assesses how effective the machine learning techniques are when applied to quantized vectors. Section 4 concludes with giving the final decision.

Vector Quantization

Vector quantization is one of the important techniques used in image processing. Carrying out this process in color reduction, image compression and segmentation makes quantization important. Octree, one of the most used quantization approaches, reduces the image to powers of 8. It is one of the image-dependent methods classified as a hierarchical clustering method. Even though its performance is low, its speed is an important advantage. Another method developed in 1980 is the LBG approach. The LBG method first determines random starting points on the input set. Then, the distances of the cluster elements to the starting points are calculated and the process continues until the optimum situation is achieved. Additionally, LBG It has an iterative structure. There is a stopping threshold constant for the number of iterations. Figure 1 shows the original pepper image, reduced to 8 colors using the octree and LBG methods.

(a) (b) (c)

Figure 1. Pepper image: a) original b) octree 8 colors c) LBG 8 colors

The loss of information in reduced images is an expected situation. The information loss of pepper images reduced to 64 colors with octree and LBG approaches is shown in Table 1. Signal noise is shown in ratio (PSNR).

Table 1. 64 color pepper image PSNR results

Method	PSNR (dB)
Octree	28.59
LBG	24.42

The PSNR is frequently used to evaluate how well an image has been reconstructed. As seen in Table 1, according to the PSNR result, octree is approximately 14.58% more successful than LBG. For this reason, while classification performances are examined in the study, only octree is used as vector quantization.

METHODS

In this study, machine learning-based classification is performed on 64 color features obtained from the Corel 1K data set using the octree method. In classification, TensorFlow 2.0 and above library and Python programming language are used. As a result of the classification, detailed findings for each approach were presented and evaluated.

At the beginning of the study, the data set was divided into training and test data. 20% of the data set is used for testing. Additionally, the standard scaling process was applied to the training data and the features were scaled to have a mean value of 0 and a standard deviation of 1. This process ensures that the data is suitable for model training.

Success and performance of classification processes, confusion. It is evaluated with the matrix technique. Indicators and mathematical models obtained from the relevant method; accuracy (1), precision (2), recall (3) and F-1 score (4) are given below:

$$\boldsymbol{Accuracy} = \frac{TN+TP}{TP+FP+TN+FN} \quad (1)$$

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\boldsymbol{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\boldsymbol{F1-Score} = \frac{2x(RecallxPrecision)}{Recall+Precision} \quad (4)$$

Figure 2 shows how success and performance evaluation are done in the classification process.

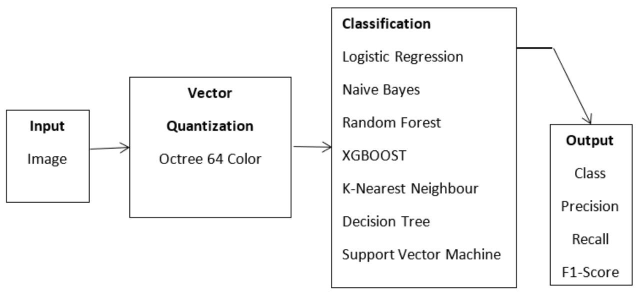

Figure 2. Block diagram of machine learning methods on quantized vectors

The average accuracy rate resulting from the classification made with the Logistic Regression algorithm was calculated at 82.50%. Figure 3 shows the confusion matrix values of the algorithm.

```
[[12  1  0  0  0  1  0  0  0  1]
 [ 1  8  1  0  0  0  0  0  1  0]
 [ 5  2 17  1  0  2  0  0  0  0]
 [ 0  1  1 20  0  0  0  0  0  0]
 [ 0  0  0  0 23  0  0  0  0  0]
 [ 0  3  0  0  0 13  0  1  1  0]
 [ 0  0  0  1  0  0 18  0  0  1]
 [ 0  1  0  0  0  1  0 22  0  0]
 [ 0  1  3  1  0  1  0  0  9  0]
 [ 1  0  0  1  0  0  0  0  0 23]]
```

Figure 3. Confusion matrix of the logistic regression algorithm

Table 2 shows the recallprecision and F1-Score indicators obtained according to the logistic regression algorithm.

Table 2. Logistic regression algorithm recall, precision and F1-Score values

Class	Precision	Recall	F1-Score
0	0.63	0.80	0.71
1	0.47	0.73	0.57
2	0.77	0.63	0.69
3	0.83	0.91	0.87
4	1.00	1.00	1.00
5	0.72	0.72	0.72
6	1.00	0.90	0.95
7	0.96	0.92	0.94
8	0.82	0.60	0.69
9	0.92	0.92	0.92

The average accuracy rate resulting from the classification made with the naive bayes algorithm was calculated at 71%. Figure 4 shows the confusion matrix values of the algorithm.

```
[[ 5  0  2  4  0  2  0  1  0  1]
 [ 1  2  5  0  0  0  0  0  2  1]
 [ 0  0 23  1  0  3  0  0  0  0]
 [ 1  1  0 19  0  0  0  1  0  0]
 [ 0  0  3  0 19  0  0  0  0  1]
 [ 0  1  2  0  0 13  0  1  1  0]
 [ 0  0  0  2  0  0 13  0  0  5]
 [ 0  0  0  0  0  0  0 24  0  0]
 [ 0  0  6  2  0  2  0  0  5  0]
 [ 4  1  0  1  0  0  0  0  0 19]]
```

Figure 4. Naive confusion about the bayes algorithm matrix

Table 3 shows the recall, precision and F1-Score indicators obtained according to the naive bayes algorithm.

Table 3. Naive bayes algorithm recall, precision and F1-Score values

Class	precision	Recall	F1-Score
0	0.45	0.33	0.38
1	0.40	0.18	0.25
2	0.56	0.85	0.68
3	0.66	0.86	0.75
4	1.00	0.83	0.90
5	0.65	0.72	0.68
6	1.00	0.65	0.79
7	0.89	1.00	0.94
8	0.62	0.33	0.43
9	0.70	0.76	0.73

The average accuracy rate resulting from the classification made with the random forest algorithm was calculated at 81.5%. Figure 5 shows the confusion matrix values of the algorithm.

```
[[11  0  0  1  0  1  0  0  0  2]
 [ 1  5  1  1  0  2  0  0  1  0]
 [ 0  2 19  1  0  3  0  0  2  0]
 [ 1  0  0 19  0  1  0  0  1  0]
 [ 0  0  0  0 23  0  0  0  0  0]
 [ 0  0  0  0  0 18  0  0  0  0]
 [ 0  0  0  0  0  0 19  0  0  1]
 [ 1  0  0  0  0  0  0 23  0  0]
 [ 0  4  1  2  0  2  0  0  6  0]
 [ 3  0  0  1  0  0  0  0  1 20]]
```

Figure 5. Confusion matrix of the random forest algorithm

Table 4 shows the recall, precision and F1-Score indicators obtained according to the random forest algorithm.

Table 4. Random forest algorithm recall, precision and F1-Score values

Class	precision	Recall	F1-Score
0	0.65	0.73	0.69
1	0.45	0.45	0.45
2	0.90	0.70	0.79
3	0.76	0.86	0.81
4	1.00	1.00	1.00
5	0.67	1.00	0.80
6	1.00	0.95	0.97
7	1.00	0.96	0.98
8	0.55	0.40	0.46
9	0.87	0.80	0.83

The average accuracy rate of the classification process has also been applied to other machine learning approaches and the results obtained are given in Table 5. Additionally, Figure 6 shows the accuracy rate graph of the applied models.

Table 5. Accuracy rates of classifiers

Class	Model	Accuracy
0	Logistic Regression	82.5
1	Naive Bayes	71.0
2	Random Forest	81.5
3	**XGBOOST**	**85.5**
4	K-Nearest Neighbour	70.5
5	Decision Tree	71.5
6	Support Vector Machine	79.0
7	stacking classification	83.0

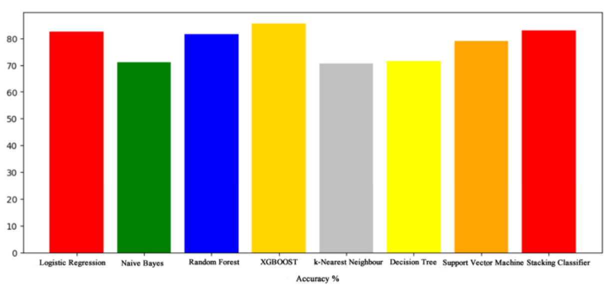

Figure 6. Accuracy rates graph of applied models

As can be seen in Table 5 and Figure 6, the XGBOOST method has the highest accuracy at 85.5%. Ranking success, from high to low, are XGBOOST, stacking classifier, logistic regression, random forest, support vector machine, decision tree, naive bayes and K-nearest neighbor.

CONCLUSION

Histograms are an important feature of images. Its simple calculation makes it a widely used identifier. However, the histogram of an image defined in the RGB color space consists of 3 different channels. Additionally, each color channel has 256 components. Performing operations on 3 different color channels and combining them is an important issue. For this reason, Corel's octree and LBG color reduction approaches images in the 1K dataset were reduced to 64 colors and a one-dimensional color histogram was obtained. According to the PSNR results, octree gave better results in color reduction. Finally, the XGBOOST method showed better performance according to the accuracy rate metric in classifying histograms with various machine learning methods.

REFERENCES

1. Lloyd, S. P. (1957). Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd, SP: Least squares quantization in PCM. *IEEE Trans. Inform. Theor.* (1957/1982), 18(11).
2. Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21, 768-769.
3. Lejeune Dirichlet, G. (1850). Over the reduction of the positive square shapes with three indefinite entire numbers . *Journal for the pure and applied Mathematics (Crelles Journal),* 1850(40), 209-227.
4. Voronoi, G. (1908). News apps of the settings continue with the theory of the shapes quadratic. Second memory. Research on _ parallelohedrons primitives . Newspaper for die queen und angewandte Mathematik *(Crelles Journal)*, 1908(134), 198-287.
5. Linde, Y., Buzo, A., & Gray, R. (1980). An algorithm for vector quantizer design. *IEEE Transactions on communications*, 28(1), 84-95.
6. Gersho, A. (1982). On the structure of vector quantizers. *IEEE Transactions on Information Theory*, 28(2), 157-166.
7. Gray, R. (1984). Vector quantization. *IEEE Assp Magazine,* 1(2), 4-29.
8. Makhoul, J., Roucos, S., & Gish, H. (1985). Vector quantization in speech coding. Proceedings of the IEEE, 73(11), 1551-1588.
9. Meagher, D. (1982). Geometric modeling using octree encoding. Computer graphics and image processing, 19(2), 129-147.
10. Uhr, L. (1963). " Pattern recognition" computers as models for form perception. Psychological Bulletin, 60(1), 40.
11. Li, D., & Wu, M. (2021). Pattern recognition receptors in health and diseases. Signal transduction and targeted therapy, 6(1), 291.
12. Serey, J., Alfaro, M., Fuertes, G., Vargas, M., Durán, C., Ternero, R., ... & Sabattin, J. (2023). Pattern recognition and deep learning technologies, enablers of industry 4.0, and their role in engineering research. *Symmetry*, 15(2), 535.
13. Mantel, N., & Brown, C. (1974). Alternative tests for comparing normal distribution parameters based on logistic regression. *Biometrics*, 485-497.
14. Kononenko, I. (1989). ID3, sequential Bayes, naive Bayes and Bayesian neural networks. Proc. of European Working Session on Learning EWSL, 4-6.
15. Palmer, E. M., & Schwenk, A. J. (1979). On the number of trees in a random forest. *Journal of Combinatorial Theory, Series B,* 27(2), 109-121.
16. Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. Annals of statistics, 1189-1232.
17. Fukunaga, K., & Narendra, P. M. (1975). A branch and bound algorithm for computing k-nearest neighbors. IEEE transactions on computers, 100(7), 750-753.
18. Magee, J. F. (1964). Decision trees for decision making (pp. 35-48). Brighton, MA, USA: Harvard Business Review.
19. Vapnik, V., Golowich, S., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. Advances in neural information processing systems, 9.

Ufuk Tanyeri

Lecturer Dr. Ufuk Tanyeri is a full-time academic staff in the department of Computer Technologies at Ankara University. He received BA (2001) degrees of Computer Education, MA (2009) degrees of Computer Education in Institute of Information at Gazi University, and PhD (2015) degree of Industrial Technological Education in Natural and Applied Sciences at Gazi University. His research interests include Distance Education, Mobil Area Adhoc Network, and Image Processing.