Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2022;14(2):81-9

ORİJİNAL ARAŞTIRMA ORIGINAL RESEARCH DOI: 10.5336/biostatic.2022-89317

Random Ferns in the Classification of High Dimension Low Sample Size Settings: A Simulation Study

Yüksek Boyut Düşük Örneklem Genişliğine Sahip Verilerin Sınıflandırılmasında Random Ferns: Bir Benzetim Çalışması

Ülger AYDOĞAN CULHA[a], Erdem KARABULUT[b]

[a]General Directorate of Turkish Highways, Ankara, Türkiye
[b]Department of Biostatistics, Hacettepe University Faculty of Medicine, Ankara, Türkiye

ABSTRACT Objective: Nowadays, "high dimension low sample size (HDLSS) settings" is very popular in many areas such as genetics, bioinformatics, medical imaging. In this study, it was aimed to investigate the classification performance of Random Ferns, which is a relatively new classifier, on HDLSS settings with different characteristics. **Material and Methods:** By simulation studies, artificial data sets that have different characteristics in terms of dimension, sample size, correlation structure, noise ratio and prevalence, were generated. Each scenario was iterated for 1,000 times and, classification performances of Random Ferns has been compared with support vector machines (SVM), which stand out with its high classification performance. **Results:** The performance variation of Random Ferns differs from SVM. It performed better at small sample size (n=20). When the F values are examined, it is seen that the classification performance of Random Ferns in the case of imbalanced distribution does not change much according to the balanced distribution situation and is higher than SVM. It has also been observed that high success as was expected in the classical data structure cannot be achieved. **Conclusion:** It is noteworthy that Random Ferns outperforms, especially in balanced distribution and small sample sizes. It is thought that this method, which does not have many applications in the field of health, will contribute to studies where the number of observations is quite low.

Keywords: Random Ferns; classification; machine learning; high dimensional data; simulation

ÖZET Amaç: Günümüzde "yüksek boyut düşük örneklem genişlikli (YBDÖG) düzenler"; genetik, biyoinformatik, tıbbi görüntüleme gibi birçok alanda oldukça popülerdir. Bu çalışmada, nispeten yeni bir sınıflandırıcı olan Random Ferns algoritmasının farklı özelliklere sahip YBDÖG yapılar üzerindeki sınıflandırma performanslarının araştırılması amaçlanmıştır. **Gereç ve Yöntemler:** Benzetim teknikleri kullanılarak boyut, örneklem genişliği, korelasyon yapısı, gürültü oranı ve prevalans açısından farklı özelliklere sahip yapay veri setleri üretilmiştir. Her bir senaryo 1.000 kez tekrarlanmış ve Random Ferns algoritmasının sınıflandırma performansı, yüksek sınıflama başarısı ile bilinen destek vektör makineleri (DVM) ile karşılaştırılmıştır. **Bulgular:** Random Ferns'in performans değişimi DVM'den farklıdır. En belirgin özelliği, küçük örneklem düzeylerinde (n=20) daha iyi performans göstermesidir. F değerleri incelendiğinde, dengesiz dağılım durumunda Random Ferns'lerin sınıflandırma performansının dengeli dağılım durumuna göre çok fazla değişmediği ve DVM'den daha yüksek olduğu görülmektedir. Klasik veri yapılarında olduğu gibi yüksek başarının ise sağlanmadığı da görülmüştür. **Sonuç:** Random Ferns'ün özellikle dengeli dağılıma ve küçük örneklem genişliğine sahip yapılarda daha iyi performans göstermesi dikkat çekmektedir. Sağlık alanında fazla uygulaması olmayan bu yöntemin, gözlem sayılarının oldukça düşük olduğu çalışmalara katkı sağlayabileceği düşünülmektedir.

Anahtar kelimeler: Random Ferns; sınıflandırma; makine öğrenmesi; yüksek boyutlu veriler; simülasyon

As Donoho stated, this century can be described as the data century.[1] In environments such as cloud, internet or electronic records systems, new information is recorded every second and huge data repositories are created. This exposes researchers to very large databases that contain a lot of information hidden in high dimensions. In order to reveal hidden information and patterns, data mining algorithms have been offered.[2,3] In parallel with the rapid development of technology, new tools are being developed to solve different

Correspondence: Ülger AYDOĞAN CULHA
General Directorate of Turkish Highways, Ankara, Türkiye
E-mail: ulgeraydogan@gmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 28 Feb 2022 ***Received in revised form:*** 06 Apr 2022 ***Accepted:*** 12 Apr 2022 ***Available online:*** 25 Apr 2022

problems in different disciplines. Such as in microarray gene expression studies, the expression levels of thousands or even tens of thousands of genes belonging to an individual are measured simultaneously and disease-related genes are determined. This allows the development of treatments for many complex diseases. However, the high cost of analyses with these tools, which we encounter in many fields such as genetics, medical imaging, proteomics, and spectroscopy, limits the researchers in the number of observations.[4,5] This leads to the emergence of data structures in which the number of observations is much lower than the number of variables as called high dimensional low sample size (HDLSS) settings.[6-8] However, there are some problems in the analysis of these settings whose importance is increasing.[9] As is known, classical multivariate statistical methods were generally developed to make an inference in data sets with a small number of well-defined variables and much more observations than the number of variables. Inability to provide the assumptions and the curse of dimensionality render classical approaches inadequate in the analysis of HDLSS settings.[10] Also, it is important to examine the performances of machine learning algorithms, which were originally developed for high-dimensions, in cases where the number of observations is low. Even though different studies have been conducted regarding the geometric structures of HDLSS settings, there is a need for comprehensive studies showing the effects of changes in the characteristics of these settings on the classification performances of current machine learning algorithms.[11,12]

In this study, it was aimed to evaluate the Random Ferns (RFerns) classifier on HDLSS settings. Within the scope of this, artificial data sets that have different characteristics in terms of dimension, sample size, correlation structure, noise ratio which is especially important in HDLSS settings, and prevalence were generated and classification performances of RFerns has been compared with support vector machines (SVM), which stand out with its high classification performance.

RANDOM FERNS

RFerns is a relatively new machine learning algorithm. It is a simple and effective algorithm developed to determine congruences in images on two screens.[13] It is especially designed to be used instead of trees for classification of parts in image recognition analysis. It is stated that the method is simpler and faster than trees and is at least as reliable as them.[14] RFerns bases its main idea on the Semi-Naive Bayes algorithm. Objects are classified with non-hierarchical structures called “ferns.” In RFerns, branching is top-down similar to Random Forest, but there is no hierarchical order among the variables used in branching. In this method, each fern contains binary test sets and the binary tests used as classifiers are chosen randomly. These tests are converted into probabilities to determine which class to assign the parts/images of interest and the results are combined with the Naive Bayes approach. In the learning phase, these sequential test sets are applied to all observations. And, leaves contain posterior probabilities. In RFerns, d variables are divided into m small sets of s size and it is assumed that these sets, called ferns, are completely independent from each other, but that the variables in the set are related. With the Semi-Naive Bayes approach, since the relationships within each fern are considered and no relationship is assumed between the ferns, both the simplicity of the standard Naive Bayes are preserved and the correlation balance is provided.[15,16]

RFerns was adapted by Kursa to work with both continuous and categorical variables as well as imbalanced distributions.[17] Thresholds values for continuous variables and sub-category groups for categorical variables are randomly determined to preserve the stochastic nature of the method. For continuous variables, the average of two random values of the relevant variable is used. In categorical variables, a category is randomly selected, except for the category that does not contain all or none of the observations in all categories of the relevant variable. It is stated that such separation protects from overfitting and reveals interactions that are not easily understood.

MATERIAL AND METHODS

In this study, artificial data sets derived by simulation techniques were used to evaluate the performance of classification methods. All analyses were performed in R software (http://www.R-project.org) utilizing the following functions: ksvm (in 'e1071' packages), and rFerns (in 'rFerns' packages). In artificial data sets, 75% of the observations are splitted as training sets and 25% as test sets. Each scenario was repeated 1,000 times, and the performances of classifiers on the test sets were reported as an average.

SCENARIOS

Considering the HDLSS settings, scenarios were created according to the sample size, the number of variables (dimensions), the relationship between dependent and independent variables, the relationship between independent variables, prevalence, and noise ratios. The general framework of the scenarios is as follows;

Sample size	• n=20, 50, 100
Number of variable (dimension)	• d=100, 500, 1000
Correlations between dependent and independent variables	• High (0.70-0.90) • Medium (0.40-0.60) • Low (0.10-0.30)
Correlations between independent variables	• High (0.70-0.90) • Medium (0.40-0.60) • Low (0.10-0.30)
Noise ratio	• High (90%), medium (50%)
Prevelances	• Balanced (P(X=1)=0.50), imbalanced (P(X=1)=0.35)

General Framework of the Scenarios

In the study, only two-class classification problems were considered. In cases where the noise ratio is determined as 50% (medium), if there is a high correlation between the dependent variable and the independent variables, half of the variables are unrelated, while the other half is highly correlated, creating a mismatch in terms of correlation levels between the independent variables and the variance covariance matrix is not positively defined. Therefore, these scenarios were excluded from the study.

DATA GENERATION PROCESSES

In the data generation process, simultaneous binary and continuous variable generation procedures and some functions in the BinNor package were used.[18-20] Only continuous independent variables were considered in the study. One of the important steps in this process is the creation of the correlation matrix. Since all the data have standard normal distribution and the dependent variable is not real discrete (dichotomous), to

create the correlation matrix, biserial correlation for the relationships between the independent continuous variable and the dependent binary variable, and the Pearson correlation coefficient for the relationships between the continuous variables were calculated by "compute.sigma.star" function. Instead of assigning a fixed value for each element in the correlation matrix, a random selection is made from the relevant range for the correlation level determined in each iteration. The inclusion of noise ratios in the scenarios was also considered during the creation of the correlation matrices. Then, data were derived from the multivariate standard normal distribution for the number of observations and the number of variables determined based on this matrix. The dependent variable was transformed to categorical according to the determined prevalence level. The data derivation function in BinNor was not used in order to ensure that prevalence remains constant at the determined level at each iteration in the creation of the dependent variable.

EVALUATION OF CLASSIFICATION PERFORMANCES

Performance criteria are of great importance in evaluating the success of learning models. Although there are many different criteria in evaluating the success of learning models, accuracy and F-measure (in cases of imbalanced distribution) were considered within the scope of this study.[21] The confusion matrix for binary classification problems is given in Table 1.

TABLE 1: Confusion matrix.

		Actual condition	
		Positive (+)	Negative (-)
Predicted Condition	Positive (+)	True positive (a)	False positive (b)
	Negative (-)	False negative (c)	True negative (d)

Accuracy is one of the most common and simple metric. It can be calculated from the confusion matrix as the sum of correct cells in the table (true positives and true negatives) divided by all cells in the table.

$$Accuracy = \frac{a+d}{a+b+c+d}$$

F-measure is the harmonic mean of precision and recall. It is a better metric when there are imbalanced classes.

$$F\ Measure = 2x\frac{precisionxrecall}{precision+recall} = \frac{2a}{2a+b+c}$$

RESULTS

In the case of balanced distribution and high noise levels, it is seen in Figure 1 that correlation levels between the independent variables do not cause a significant change in SVM performance. In parallel with the literature, performance increases as the number of observations increases, while performance decreases as the dimensions increase. It is noteworthy that when the number of dimensions is 500 and 1,000, the results are closer to each other. In the case of d=100, SVM is more affected by the relationship between dependent and independent variables. The performance variation of RFerns differs from SVM (Figure 1). The most prominent feature is that RFerns performed better at low number of observations (n=20). When the relationship between the dependent and the independent variables increases, the performance of RFerns also improves at low number of dimensions.

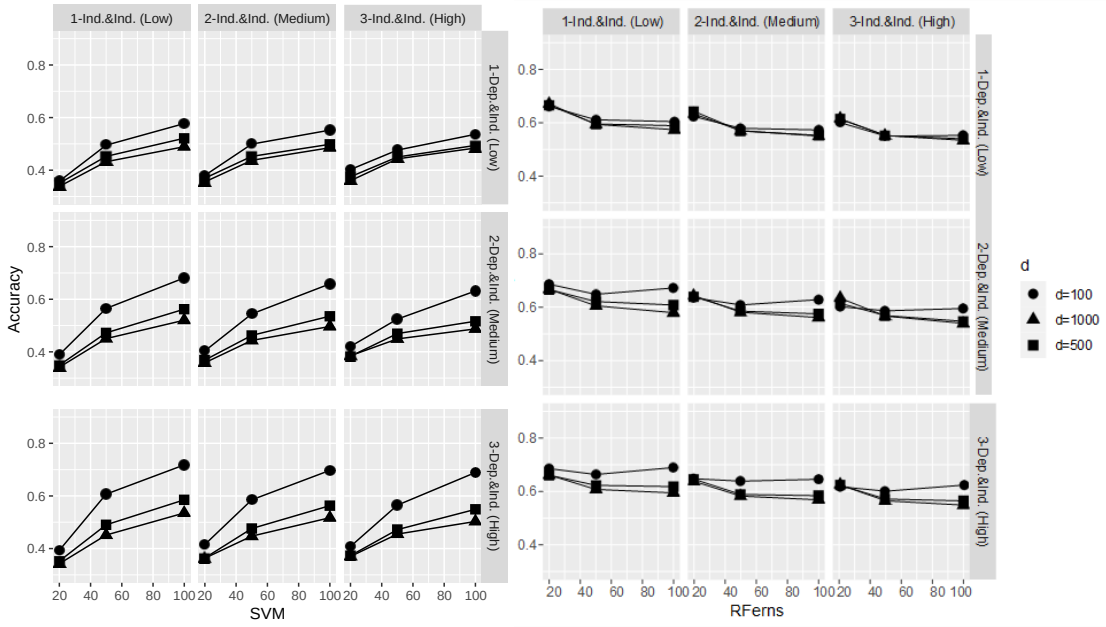

FIGURE 1: Classification accurary of SVM & RFerns (balanced distribution with high noise).
SVM: Support vector machines.

In the presence of medium noise (50%) and the balanced distribution, classification performances of the two algorithms are shown in Figure 2. It is seen that the classification performance of SVM is generally improved in case of the medium noise level. RFerns shows a different tendency compared to SVM in the conditions determined for balanced distribution with medium noise level as well as in the presence of high noise level. Contrary to the high noise level, it is seen that the increase in the correlation between the independent variables creates a slightly more pronounced effect here.

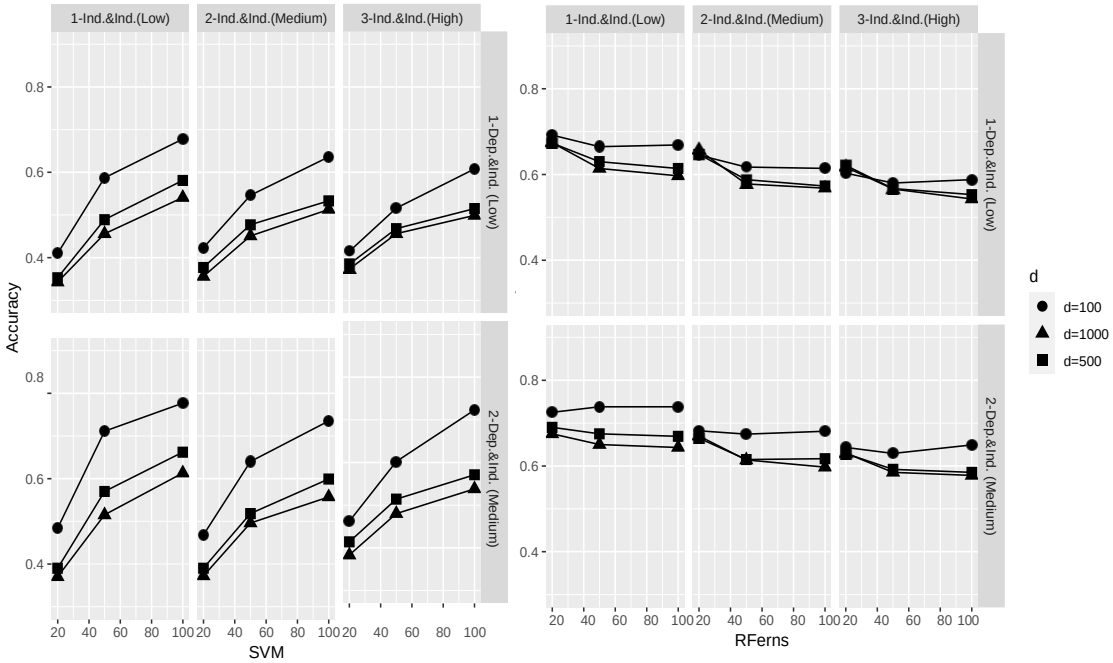

FIGURE 2: Classification accuracy of SVM & RFerns (balanced distribution with medium noise).
SVM: Support vector machines.

The results for the scenarios created for imbalanced distribution and high noise level are given in Table 2, and the results for the scenarios created for imbalanced distribution and medium noise level are given in Table 3.

When the F values are examined, it is seen that the classification performance of the SVM decreases considerably in the case of imbalanced distribution, contrary to what is observed in the accuracy values. On the other hand, it is seen that the classification performance of RFerns does not change much according to the balanced distribution situation and the F values are higher than SVM.

TABLE 2: Classification performances (imbalanced distribution with high noise).

Sample size (n)	Dimesion (d)			Low correlation (Dep.&Ind.)			Medium correlation (Dep.&Ind.)			High correlation (Dep.&Ind.)		
				Low Corr. (Ind.& Ind.)	Medium Corr. (Ind.& Ind.)	High Corr. (Ind.& Ind.)	Low Corr. (Ind.& Ind.)	Medium Corr. (Ind.& Ind.)	High Corr. (Ind.& Ind.)	Low Corr. (Ind.& Ind.)	Medium Corr. (Ind.& Ind.)	High Corr. (Ind.& Ind.)
20	100	SVM	Acc.	0.63	0.624	0.607	0.645	0.617	0.614	0.634	0.638	0.607
			F Mea.									
		Rferns	Acc.	0.374	0.394	0.437	0.368	0.405	0.433	0.376	0.398	0.444
			F Mea.	0.498			0.486			0.494	0.477	
	500	SVM	Acc.	0.648	0.643	0.621	0.647	0.628	0.61	0.646	0.631	0.617
			F Mea.									
		Rferns	Acc.	0.353	0.369	0.397	0.356	0.378	0.407	0.369	0.379	0.397
			F Mea.	0.486			0.491					
	1,000	SVM	Acc.	0.648	0.625	0.63	0.641	0.623	0.628	0.643	0.648	0.627
			F Mea.									
		Rferns	Acc.	0.358	0.383	0.388	0.363	0.383	0.384	0.36	0.362	0.387
			F Mea.	0.485			0.498			0.493	0.482	0.488
50	100	SVM	Acc.	0.631	0.629	0.627	0.644	0.63	0.63	0.653	0.644	0.641
			F Mea.	0.038	0.052	0.055	0.07	0.098	0.078	0.102	0.128	0.131
		Rferns	Acc.	0.388	0.411	0.439	0.382	0.436	0.46	0.397	0.441	0.47
			F Mea.	0.525	0.508	0.486	0.52	0.526	0.507	0.523	0.523	0.51
	500	SVM	Acc.	0.64	0.63	0.623	0.635	0.628	0.633	0.649	0.628	0.631
			F Mea.	0.012	0.02	0.044			0.046	0.123	0.028	0.046
		Rferns	Acc.	0.365	0.382	0.417	0.372	0.395	0.412	0.409	0.399	0.416
			F Mea.	0.517	0.509	0.506	0.521	0.518	0.499	0.53	0.522	0.506
	1,000	SVM	Acc.	0.638	0.634	0.626	0.639	0.638	0.627	0.643	0.639	0.628
			F Mea.		0.019	0.024	0.006		0.027		0.023	
		Rferns	Acc.	0.365	0.382	0.405	0.364	0.378	0.402	0.362	0.378	0.404
			F Mea.	0.519	0.515	0.517	0.518	0.515	0.514	0.515	0.513	0.51
100	100	SVM	Acc.	0.652	0.647	0.647	0.673	0.674	0.667	0.693	0.686	0.681
			F Mea.	0.068	0.054	0.056	0.186	0.186	0.149	0.272	0.256	0.238
		Rferns	Acc.	0.397	0.428	0.457	0.424	0.461	0.486	0.445	0.48	0.503
			F Mea.	0.519	0.502	0.483	0.534	0.518	0.509	0.543	0.539	0.518
	500	SVM	Acc.	0.649	0.646	0.643	0.649	0.649	0.642	0.681	0.651	0.648
			F Mea.	0.013	0.021	0.024	0.021	0.025	0.032	0.248	0.039	0.034
		Rferns	Acc.	0.366	0.395	0.417	0.373	0.399	0.431	0.478	0.41	0.433
			F Mea.	0.514	0.505	0.495	0.518	0.508	0.507	0.553	0.514	0.506
	1,000	SVM	Acc.	0.645	0.645	0.647	0.646	0.648	0.643	0.654	0.649	0.644
			F Mea.	0.007	0.014	0.019	0.01	0.02	0.019	0.013	0.017	0.025
		Rferns	Acc.	0.363	0.384	0.404	0.364	0.386	0.413	0.359	0.387	0.413
			F Mea.	0.517	0.511	0.495	0.517	0.509	0.507	0.509	0.508	0.505

SVM: Support vector machines.

TABLE 3: Classification performances (imbalanced distribution with medium noise).

Sample size (n)				20			50			100		
Dimension (d)				100	500	1000	100	500	1000	100	500	1000
Low correlation (Dep.&Ind.)	Low correlation (Ind.&Ind.)	SVM	Acc.	0.631	0.644	0.643	0.645	0.639	0.637	0.671	0.651	0.646
			F Mea.				0.097	0.027	0.015	0.181	0.059	0.026
		Rferns	Acc.	0.381	0.362	0.359	0.396	0.377	0.369	0.451	0.397	0.38
			F Mea.	0.497	0.489	0.495	0.524	0.52	0.521	0.536	0.524	0.521
	Medium correlation (Ind.&Ind.)	SVM	Acc.	0.616	0.634	0.639	0.63	0.632	0.628	0.657	0.647	0.651
			F Mea.				0.09	0.045	0.027	0.143	0.053	0.038
		Rferns	Acc.	0.41	0.382	0.366	0.444	0.403	0.394	0.482	0.43	0.407
			F Mea.		0.493		0.526	0.518	0.524	0.53	0.516	0.509
	High correlation (Ind.&Ind.)	SVM	Acc.	0.615	0.619	0.622	0.627	0.625	0.629	0.65	0.644	0.646
			F Mea.				0.11	0.048	0.031	0.127	0.044	0.035
		Rferns	Acc.	0.437	0.408	0.394	0.474	0.42	0.407	0.491	0.44	0.421
			F Mea.			0.498	0.511	0.504	0.515	0.513	0.5	0.503
Medium correlation (Dep.&Ind.)	Low correlation (Ind.&Ind.)	SVM	Acc.	0.65	0.643	0.635	0.684	0.643	0.638	0.755	0.672	0.654
			F Mea.				0.294	0.093	0.049	0.527	0.193	0.084
		Rferns	Acc.	0.381	0.366	0.367	0.464	0.397	0.379	0.572	0.456	0.41
			F Mea.		0.487	0.498	0.555	0.528	0.522	0.601	0.547	0.531
	Medium correlation (Ind.&Ind.)	SVM	Acc.	0.643	0.633	0.629	0.659	0.632	0.633	0.717	0.657	0.651
			F Mea.				0.234	0.082	0.047	0.415	0.118	0.058
		Rferns	Acc.	0.428	0.393	0.389	0.51	0.431	0.404	0.572	0.479	0.438
			F Mea.				0.551	0.529	0.52	0.573	0.529	0.52
	High correlation (Ind.&Ind.)	SVM	Acc.	0.618	0.63	0.623	0.648	0.627	0.628	0.706	0.648	0.648
			F Mea.					0.091	0.049	0.353	0.089	0.051
		Rferns	Acc.	0.471	0.42	0.398	0.526	0.456	0.425	0.567	0.483	0.453
			F Mea.				0.545	0.519	0.517	0.555	0.519	0.515

SVM: Support vector machines.

DISCUSSION

It is thought that the fact that HDLSS settings are different from classical structures is generally ignored. The number of observations and dimensions, the relations between dependent variable and independent variables are kept in the foreground, but the presence of noise variables, which is one of the most important problems encountered in microarray data, is mostly neglected in studies conducted with simulation applications on HDLSS structures in the literature.[22-25] In this study, the effects of this neglected situation in practice have been tried to be observed. In order to form a step towards further studies, comprehensive scenarios were designed and multi-dimensional comparisons were tried to be made. Therefore, this study differs from others in that it considers many factors that determine data structures.

When the classification performances are carefully examined, it is seen that SVM method, which is very successful and has high classification performance in classical structures, cannot be as successful as expected in HDLSS settings. The fact that this method is frequently preferred especially in microarray studies requires careful consideration in this sense.

Considering the F measure, when the change in classification performance in the case of balanced distribution and imbalanced distribution is examined, it is seen that SVM is most affected by the prevalence change. RFerns is least affected by this situation. However, in the case of imbalanced distribution, the fact that the classification performances are mostly below 0.50 makes us think that these problems should be solved with different approaches.

Although many scenarios have been studied here, it is thought that this scope should be expanded with further studies in order to reach more general conclusions. It will be beneficial to reveal structures more similar to microarray data, developing scenarios where the number of dimensions is over tens of thousands. In addition to the dimension, it is also important to repeat the studies with different observation numbers and different noise levels within the scope of the HDLSS settings. Another important point is that the model parameters, which are default in R software, were used within the scope of this study. It is thought that different results can be obtained in this regard with further studies to be carried out by considering the model parameters.

CONCLUSION

The results of the study reveal that the classification performances of machine learning algorithms, which are frequently used in the literature and especially encountered in classification problems with microarray data, show different tendencies in different HDLSS data structures. When the classification performances are carefully examined, it is seen that the algorithms in question -especially SVM- do not have high performance. However, it is noteworthy that RFerns method outperforms the others, especially in structures with balanced distribution and low sample size. It is thought that this method, which does not have many applications in the field of health, will contribute to microarray and medical imaging studies where the number of observations is quite low.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Ülger Aydoğan Culha, Erdem Karabulut;* ***Design:*** *Ülger Aydoğan Culha, Erdem Karabulut;* ***Control/Supervision:*** *Ülger Aydoğan Culha, Erdem Karabulut;* ***Data Collection and/or Processing:*** *Ülger Aydoğan Culha;* ***Analysis and/or Interpretation:*** *Ülger Aydoğan Culha, Erdem Karabulut;* ***Literature Review:*** *Ülger Aydoğan Culha;* ***Writing the Article:*** *Ülger Aydoğan Culha;* ***Critical Review:*** *Erdem Karabulut.*

REFERENCES

1. Donoho DL. High-dimensional data analysis: the curses and blessings of dimensionality. AMS Lectures. 2000. [Link]
2. Tan PN, Steinbach M, Kumar V. Introduction to Data Mining. 1st ed. New York: Pearson Education; 2006.
3. Han J, Kamber M. Data Mining: Concept and Techniques. 2nd ed. Amsterdam ; Boston: Elsevier Inc.; 2006.
4. Qiao Z, Zhou L, Huang JZ. Effective linear discriminant analysis for high dimensional, low sample size data. Proceedings of the World Congress of Engineering. 2008;1070-5.
5. Ramey JA, Young PD. A comparison of regularization methods applied to the linear discriminant function with high-dimensional microarray data. Journal of Statistical Computation and Simulation. 2011;83(3):581-96. [Crossref]
6. Ye J, Wang T. Regularized discriminant analysis for high dimensional, low sample size data. Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006. p.454-63. [Crossref] [PubMed]
7. Sen PK, Tsai MT, Jou YS. High dimension, low-sample size perspectives in constrained statistical inference: The SARSCoV RNA genome illustration. Journal of the American Statistical Association. 2007;102(478):686-94. [Crossref]
8. Muller KE, Chi Y, Ahn J, Marron JS. Limitations of high dimension, low sample size principal components for gaussian data. 2008. [Link]
9. Yang X. Machine Learning Methods in HDLSS Settings. Chapel Hill: North Carolina University; 2021. [Link]
10. Qiao X, Zhang HH, Liu Y, Todd MJ, Marron JS. Asymptotic properties of distance-weighted discrimination. Journal of the American Statistical Association 2010;105:401-14. [Crossref] [PubMed] [PMC]
11. Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. J R Statist Soc Series B. 2005;67:476-98. [Crossref]

12. Ahn J. High Dimension Low Sample Size Data Analysis. Chapel Hill: North Carolina University; 2006. [Link]
13. Özuysal M, Calonder M, Lepetit V, Fua P. Fast keypoint recognition using random ferns. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010;32(3):448-61. [Crossref] [PubMed]
14. Bosch A, Zisserman A, Munoz X. Image Classification Using Random Forests and Ferns. IEEE 11th International Conference on Computer Vision; Oct 14-21, 2007. p.1-8. [Crossref]
15. Dong Y, Zhang Y, Yue J, Hu Z. Comparison of random forest, random ferns and support vector machine for eye state classification. Multimed Tools Appl 2016;75(19):11763-83. [Crossref]
16. Kim S, Ko BC. Building Deep Random Ferns Without Backpropagation. Vol 8. IEEE Access; 2020. p.8533-42. [Crossref]
17. Kursa MB. rFerns-Random Ferns Method Implementation for the General Purpose Machine Learning. Journal of Statistical Software. 2014;61(10):1-13. [Crossref]
18. Demirtas H, Doganay B. Simultaneous generation of binary and normal data with specified marginal and association structures. J Biopharm Stat. 2012;22(2):223-36. [Crossref] [PubMed]
19. Demirtaş H, Amatya A, Doğanay B. BinNor: An R package for concurrent generation of binary and normal data. Communication in Statistics-Simulation and Computation. 2014;43(3):569-79. [Crossref]
20. Demirtaş H, Amatya A. Package 'BinNor', simultaneous generation of multivariate binary and normal variables. 2012. Available from: [Link]
21. Straube S, Krell MM. How to evaluate an agent's behavior to infrequent events? Reliable performance estimation insensitive to class distribution. Frontiers in Computational Neuroscience. 2014;8(43). [Crossref] [PubMed] [PMC]
22. Cho SB, Won HH. Machine learning in DNA microarray analysis for cancer classification. In proceedings of the first Asia-Pasific bioinformatics conference on bioinformatics 2003. Volume 19 (APBC '03). Australian Computer Society, Inc., AUS; 2003. p.189-98.
23. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics. 2000;16(10):906-14. [Crossref] [PubMed]
24. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics. 2008;9 Suppl 1(Suppl 1):S13. [Crossref] [PubMed] [PMC]
25. Tamatani M. Asymptotic Theory for Discriminant Analysis in High Dimension Low Sample Size. Memoirs of the Graduate School of Science and Engineering. Series B: Mathematics. 2015;48:15-26. [Link]