Journal of Sciences, Islamic Republic of Iran 35(1): 55 - 62 (2024) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Supervised Clustering of Persian Handwritten Images

S. Moradnia, M. Golalizadeh*

Department of Statistics, Tarbiat Modares University, Tehran, Islamic Republic of Iran

Received: 21 June 2023 / Revised: 13 January 2024 / Accepted: 30 April 2024

Abstract

Clustering, a fundamental multivariate statistical method, serves as a valuable tool for extracting meaningful insights from complex datasets. Analyzing high-dimensional data, however, presents challenges, notably the curse of dimensionality. While various methods have been developed to address the dimensionality reduction, most overlooked the role of dependent variables. In contrast, supervised clustering leverages the inherent information in response variables, offering substantial benefits in data dimension reduction and accelerating clustering computations. This paper evaluates the efficacy of supervised clustering in the analysis of Persian handwritten images. Focusing on the multi-class nature of Persian handwritten data, the identification of important variables for each digit not only reduces data dimensionality but also this reduction in dimensionality does not compromise the accuracy of predicting new observations at any stage of the algorithm. Additionally, the approach demonstrates relatively high accuracy in predicting the response variable. This study contributes a novel perspective toward clustering methods, highlighting the integration of supervised techniques for improved performance in high-dimensional data analysis.

Keywords: Supervised Clustering; High Dimensional Data; Dimension Reduction; Persian Handwritten Images; Hoda.

Introduction

The appearance of high-dimensional data in most scientific fields has changed statistical thinking in various related topics. To confine to the statistical analysis of high-dimensional data, the large number of predictor variables causes many problems in the analysis, interpretation and application of some statistical methods. Today, multivariate statistical methods leading to the discovery of useful information from databases and in which data analysis is done based on a large number of variables for each observation have attracted the attention of many analysts (1).

In the meantime, clustering is one of the most effective and practical multivariate techniques; its growing application across a wide range of scientific domains attests to its high demand for use. As per (2), clustering is an unsupervised technique that finds comparable data groupings without the need for prior knowledge of the corresponding groups. Clustering tasks can be performed with a wide variety of algorithms, many of which share a common implementation format. Nonetheless, these techniques differ in how they determine the labels for the objects within each cluster and how to measure similarity or distance. Having an understanding of clustering algorithms and their application might occasionally assist the researcher in selecting the approach that best fits her data. Despite the

* Corresponding Author: Tel: +98 21-82884705; Email: golalizadeh@modares.ac.ir

fact that the goal of clustering analysis is grouping observations, researchers frequently attempt to cluster variables instead. A simple illustration pertains to the analysis of high-dimensional data (3).

The process of clustering variables involves grouping them into homogeneous clusters so that each group's variables have comparable data and a high correlation with one another (4). This method can be helpful for dimension reduction tasks that target variable selection, allowing cluster indices—such as the representative of each cluster—to be used for explaining the majority of changes in the data rather than a collection of variables. However, one of the main concerns when grouping variables is whether or not there is a relationship between the response and predictor variables, meaning that in the case this relationship exists, then the variables are grouped. One of the reasons this strategy is set in the field of supervised methods is that it uses response variable information in clustering variables (5).

One technique for grouping variables based on response variable information is supervised clustering. This kind of clustering was first used in (6) to analyze gene data and identify the type of cancer tissue. This supervised method has its roots in the study described in (7), where numerous potential clusters were first produced using hierarchical unsupervised clustering as a two-step method. Then, just a few sets of variables that held the most important information for separation were found, the average of each cluster was taken into consideration as an input variable for the model. This kind of clustering is classified as a supervised algorithm because of the second stage, which depends on more data regarding the response variable.

In subsequent years, the field of supervised clustering has seen notable advancements (15) combined clustering with logistic regression, introducing a powerful approach known as “Pelora” that employs the penalized negative log-likelihood function based on the L_2 norm. Moreover, recent research efforts have extended the application of supervised clustering, incorporating sophisticated techniques such as neural networks. For instance, (19) explored the potential of neural networks in enhancing supervised clustering methodologies. In parallel, the concept of “semi-supervised clustering” has gained momentum in recent years. Researchers, such as (20), have ventured into semi-supervised clustering, specifically in the context of multi-ant colonies consensus clustering. These developments underscore the dynamic evolution of supervised clustering techniques, incorporating diverse approaches ranging from logistic regression to advanced neural network architectures.

This paper deals with the application of supervised clustering in the analysis of Persian handwritten images. For this purpose, the theoretical description of supervised clustering comes in next section. Then, the application of the method on a real example is appeared in section 2. Additionally, section 3 has been introduced to present a comparative analysis with other supervised methods. The paper ends with a general discussion and conclusion.

Materials and Methods

We provide manin materials to represent our method here. In includes the theoretical aspects of the supervised clustering, score and margin functions and ingeredients for real applications.

1. Supervised Clustering

This section presents a supervised algorithm of similarities and interactions among predictor variables for clustering in high-dimensional spaces. The basic stochastic model for high-dimensional data with response variable is assumed as random pairs $(\mathbf{X}, Y)$ with values from the multiplicative space $\mathbb{R}^p \times \mathbb{Y}$, where $\mathbf{X} \in \mathbb{R}^p$ denotes a data matrix standardized to mean zero and unit variance. Note that Y is the associated response variable, taking numeric values from the set $\mathbb{Y} = \{0,1, \dots K-1\}$, where K represents the number of states of the response variable. For the simplicity of the discussion and easy understanding of the algorithm, first the case $K = 2$ is considered. Suppose, only a few predictor variables determine the association with the response variable. Then, the conditional probability can be defined as

$$Pr(Y = 1|\mathbf{X}) = f(\tilde{\mathbf{X}}) = f\left(\tilde{X}_{C_1}, \tilde{X}_{C_2}, \dots, \tilde{X}_{C_q}\right), \qquad (1)$$

where $f(.)$ is a nonlinear function mapping from $\mathbb{R}^q$ to $[0,1]$. Also, $\{C_1, \dots, C_q\}$ are clusters of variables which $\{\cup_{i=1}^{q} C_i\} \subset \{1, \dots, p\}$ and for any $i \neq j$, $C_i \cap C_j = 0$. In addition, it is assumed that $\tilde{X}_{C_i} \in \mathbb{R}$ denotes a “representative” of each cluster C_i. Since the researcher is looking for clusters with similar variables, a simple linear combination

$$\tilde{X}_{C_i} = \frac{1}{|C_i|} \sum_{g \in C_i} \alpha_g X_g \qquad (2)$$

where $|C_i|$ is the number of variables gathered in each cluster C_i and $\alpha_g \in \{-1,1\}$, is a good option for $\tilde{X}_{C_i}$. The advantage of using this simple linear combination as a representative of each cluster is that it biases the model towards having correlated sets of variables and as a result, the speed of overfitting of the model decreases. However, finding a subset, say q, from p variables and forming clusters $\{C_1, \dots, C_q\}$ with probabilistic structure is somehow difficult. Thus, it seems necessary to provide a computationally intensive procedure that approximately solves (1) and brings reasonable results from an experimental point of view. To do so, we need the score and margin functions, both described in turn.

1.1. Score Function

Suppose $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_j \in \mathbb{R}^p$ and $y_j \in \{0,1\}$, are n independent and identically distributed $(i.i.d)$ realizations of the random vector $(\mathbf{X}, Y)$, whose the values of $\mathbf{x}_j$ are centered to mean zero and scaled to unit variance. Note that the goal is to find subsets of variables with accurate separation in binary problems. To achieve this, the Wilcoxon's test statistic, as a nonparametric rank-based score function, was proposed in (6) and is called the criterion "Wilma". Following them, the score of the i-th variable from the n-dimensional vector of the observed values of variables $\xi_i = (x_{i1}, \dots, x_{in})$, can be calculated as

$$Score(\xi_i) = s(\xi_i) = \sum_{j\in N_0} \sum_{l\in N_1} 1_{[x_{ij} \geq x_{il}]}, \qquad (3)$$

where x_{ij} is the amount of information of the i-th variable for j-th response and also N_k for $k \in \{0,1\}$ represents a subset of $\{1, \dots, n\}$. Considering the score function as the Wilcoxon's test statistic, it is possible to order variables and clusters according to their potential significance for the separation response variable. In the case that exact separation occurs, the value of the score function will take its lowest value ($s_{\min} = 0$) and otherwise, its maximum value ($s_{\max} = n_0 n_1$).

We know that the best way to separate variables is to place variables with low information load in class 0 and high information load in class 1. However, in some cases, this separation may not happen well. Therefore, as suggested in (6), in order to achieve the best separation of variables, a sign flipping of the value of each variable vector $\xi_i = (x_{i1}, \dots x_{in})$ for all the variables $i \in \{1, \dots, p\}$ can be utilized having score $s(\xi_i) > \frac{s_{\max}}{2}$, by multiplying it with (-1). In other words, considering $\alpha_g \in \{-1,1\}$, the information of the explanatory variables about the response variable can be written as a conditional relation, represented as

$$\tilde{\xi}_i = \alpha_i \xi_i = \begin{cases} \xi_i & s(\xi_i) \leq \dfrac{s_{\max}}{2} \\ -\xi_i & s(\xi_i) > \dfrac{s_{\max}}{2}. \end{cases}$$

After the sign-flip for candid variables, the scores of all variables in data matrix are equal to

$$s(\tilde{\xi}_i) = \min\{s(\xi_i), s_{\max} - s(\xi_i)\}.$$

1.2. Margin Function

Although the modified score function has many features, it does not perform well in some cases. Therefore, it is necessary to reform the score function. For this purpose, the margin function was introduced by (6) as

$$Margin(\xi_i) = m(\xi_i) = \min_{l\in N_1}(x_{il}) - \max_{j\in N_0}(x_{ij}), \qquad (4)$$

which is a continuous scale for separating the response variable and N_0, N_1 and x_{ij} are the same quantities as defined in (3). It can be seen that the value of margin function will be positive if and only if the score function is zero and $\tilde{\xi}_i$ completely separates response variable in the best way. Otherwise, the value of margin function will be negative.

In the continuation of the research, the researcher may encounter cases in which the response variable is not two-classes. Particularly, in multi-class classification problems, the goal is to select the appropriate label from $K > 2$ (the number of types of the response variable). Then, the question that arises is, "How can the supervised clustering algorithms be implemented on the multiclasses data sets? "

There are different methods to transform problems with K classes specified labels from the set $\mathbb{Y} \in \{0, \dots, K-1\}$ to problems of two classes. In this regard, a comprehensive study was conducted in (8).

Let us assume the objective is to consider a binary problem for each K class. According to this method, called "one against all", for every $k \in \{0, \dots, K-1\}$, samples with the label $\mathbb{Y} = k$ are considered as the members of the k-th class and otherwise they do not belong to the corresponding class. In this case, there will be K modes to make a decision, so that the best decision is finally made about assigning each sample to the cluster. The following section describes an implementation of supervised clustering for the analysis of Persian handwritten images, which is a multi-class dataset.

2. Real Example Analysis

Researchers in some scientific disciplines are interested in the recognition of handwritten characters. This skill, which was first made possible by databases, is crucial to image processing, handwriting recognition, zip code recognition, and other applications. Numerous researchers have worked on this topic in the last few decades. Readers interested in learning more should consult the reports in (9), (10), and (11) for additional insights.

Standard databases like "CENPARMI" and "CEDAR" are available in English for certain scripts (see, e.g. (12)). On the other hand, there are extremely few standard datasets of handwritten numbers or letters in free form for other writings, such as Farsi (Persian) scripts which writes its digits from left to right. To conduct our analysis in this article, we use a popular Persian database called "Hoda". A typical example of handwritten Farsi digits from this database is displayed

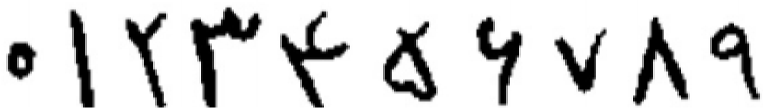


Figure 1. A typical sample of ten Farsi handwritten digits.

in Figure 1.

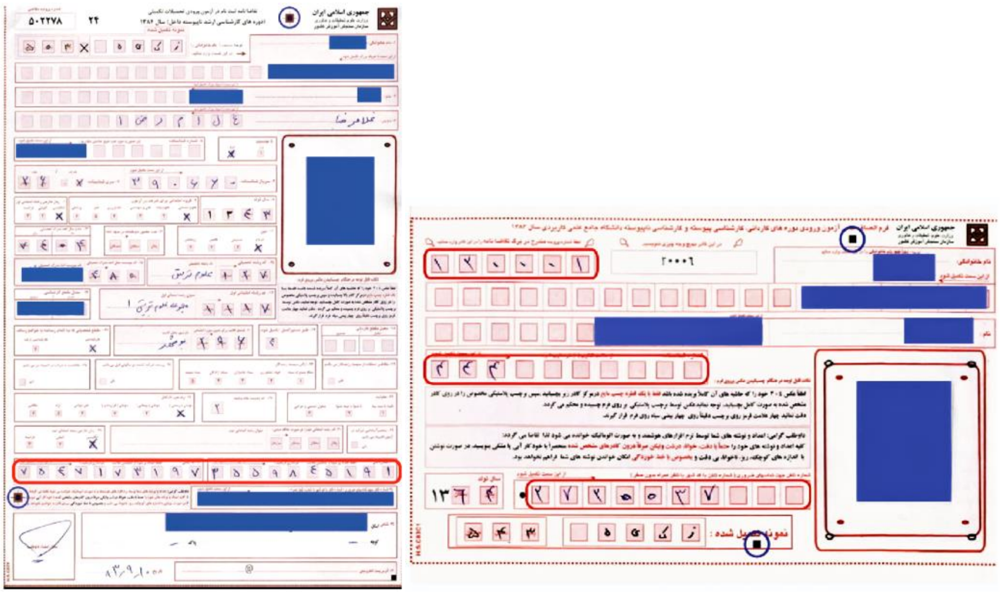
There are 102353 examples of black and white manuscripts in the "Hoda" handwritten digit collection, which is the first sizable collection of handwritten Farsi digits. This compilation was created as part of a master's thesis on handwritten form recognition. This collection's data was taken from 11942 registration forms of two different kinds. One was for the University of Applied Sciences continuous associate exam (type 2), and the other was for the national master's exam (type 1). Figure 2 displays two sample forms together with their desired fields.

5393 type 1 forms and 6549 type 2 forms were looked at in the process of gathering this data collection. It should be mentioned that all of these forms were scanned using a high-speed Axiome 4300 scanner at 200 dpi (200 dpi) in 24-bit color format. Of them, the form type 1 contains two digit fields each having ten digits: "Postal Code" and "National Code". Three digit fields, "Record Number", "Identity Certificate Number", and "Phone Number", with a maximum of 26 digits, were chosen in the form type 2. Black and blue handwritten text appears on both forms, which are colored. Following the completion of all forms, 94530 type 1 and 128203 type 2 digits were acquired. Ultimately, 60,000 digits were chosen at random from every form. The interested readers may get more information about this dataset at (13).

Results

Now, we will use a one-against-all approach to apply Wilma's algorithm to Hoda's multi-class dataset and analyze the outcomes. We treat the first 500 samples as selected samples in the first phase. Notably, each number's image is composed of a dimensional matrix; as a result, the data set we chose has 1024 variables. A user-written example of the number "Six" is displayed in Figure 3. Additionally, Figure 4 displays the dimensional matrix that forms the numeral "Six".

Using the "gap statistic" and (14), the optimal number of clusters for clustering the variables of the Hoda data set was obtained as 10 clusters. This is shown in Figure 5.



(A) (B)

Figure 2. A sample of each form filed by student: type 1 (A) and type 2 (B).

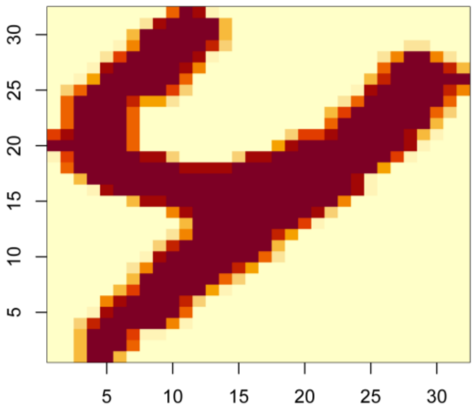


Figure 3. Number “Six” written by the user.

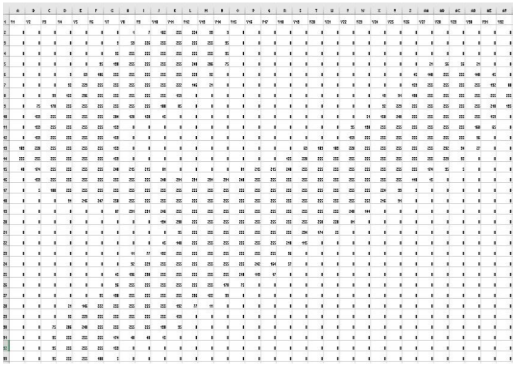


Figure 4. The digit “Six” appeared as a shadow in a matrix representation.

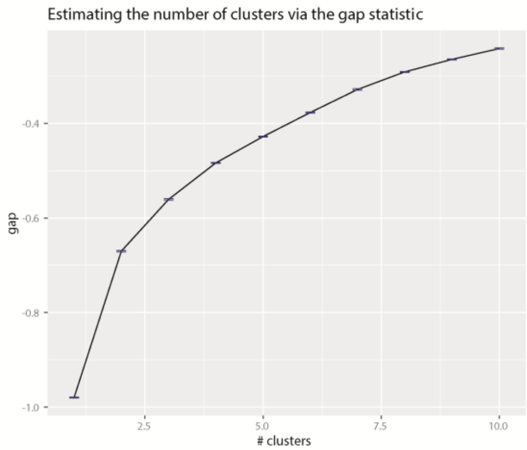


Figure 5. Estimation optimal number of clusters using gap statistic for the Hoda data set.

We used Hoda's chosen dataset and the one-against-all strategy to run the Wilma algorithm across 10 clusters in order to identify the important variables for each digit.

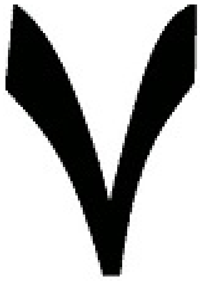


Figure 6. Digit “Seven” written in the Persian script.

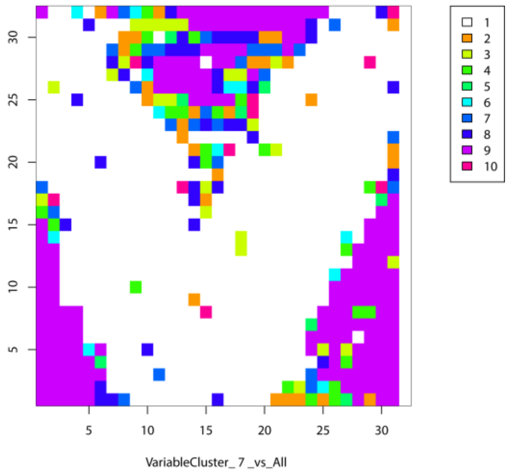


Figure 7. Important variables for the Persian digit “Seven” appeared as image.

For instance, by comparing the digit “Seven” to every other digit, Table 1 displays the important variables that Wilma's supervised algorithm detected. Figure 6 displays the written form of the Persian numeral “Seven” so that you may get an idea of its structure.

Table 1 shows that 422 out of 1024 factors were deemed significant by Wilma's method, which had a %97 accuracy rate, in order to predict the number “Seven”. The key characteristics of the digit “Seven” are displayed in Figure 7 in comparison to other digits and their clusters according to Table 1.

Figure 7 shows that these colored dots, though in Persian, stand in for significant aspects of the digit “Seven”, serving as both a template and an overall representation of the digit for the reader. Table 2 presents the predictions made by Wilma's algorithm on the Hoda data set for each digit in comparison to the others. As can be observed, the accuracy for each digit is relatively high, which makes performance of the proposed method exceptionally well.

The nature of the clustering method is commonly based on unsupervised algorithm. However, in this research, we dealt with in a different view, namely the supervised learning. Therefore, we have to compare our

Table 1. The result of clustering the digit "Seven" against others in Hoda dataset. C1 to C10 represent Clusters 1 to 10, respectively.

C1	**C2**	**C3**	**C4**	**C5**		**C6**	**C7**	**C8**						**C9**	**C10**
365	1017	8	1015	254	986	1000	115	702	608	49	834	954	1022	243	751
5	1014	9	7	713	984	73	106	703	577	46	833	958	995	783	364
860	1013	922	545	372	985	55	827	704	576	50	832	962	1023	370	604
38	6	952	178	369	1018	211	510	705	575	51	831	902	987	494	465
39	10	953	180	30	1016	210	271	706	637	52	830	901	1024	944	814
40	11	966	176	160	798	48	242	730	173	139	829	900	1019	390	546
431	416	54	202	495	170	47	882	731	543	825	828	899	991	811	
337	415	80	74	303	171	448	871	732	674	140	826	898	996	630	
336	146	78	479	305	301	589	27	733	673	605	961	897	1021	1009	
335	145	77	268	37	45	969	41	734	672	606	964	896	990	195	
334	147	447	578	306	44	950	116	735	671	607	959	895	994	134	
967	148	366	824	400	43	919	212	736	670	112	960	894	998	597	
974	149	481		782	921	76	172	737	669	111	965	893	1020	31	
629	150	982		422	923	611	270	738	206	110	1	892	988	102	
1006	273	209		462	42	610	914	762	235	109	638	891	989	332	
721	432	352		745	1005	603	401	763	641	542	639	890	992	157	
951	138	399		340	968	368	168	764	141	107	640	889	993		
70	267	513		717	1001	477		765	79	105	12	888	997		
283	63	274		300	179	229		766	113	544	13	870	999		
920	72	983		383	175	688		767	118	574	14	869	963		
143	177			103	174	304		768	142	573	24	868	957		
162	104			398	181	430		769	117	480	86	867	924		
326	333			625	88	482		770	205	668	511	866	929		
71	509			464	85	1012		771	114	53	512	865	927		
75	463			188	84	351		772	208	636	87	864	925		
	750			367	83	201		793	207	203	23	863	956		
	307			135	82	514		794	108	237	22	862	926		
	910			144	81	338		795	269	236	21	861	955		
	455			384	272	810		796	204	241	20	859	934		
				275	527	788		797	25	240	19	858	933		
				449	302	136		799	698	239	18	857	932		
				524				800	699	238	17	856	931		
				223				801	700	609	16	836	930		
				69				802	701	642	15	835	928		
								804	803						

proposed supervised clustering algorithm with other known (supervised) methods in this field rather than standard clustering methods. Hence, we will give a brief description of several supervised procedures studied in this research including the "Decision Tree", "Neural Network" and "Suport Vector Machine" algorithm. Then, we report the results of applying these techniques on Hoda dataset and compare the outcomes with our proposed method.

A "Decision Tree" which is abbreviated as "DT", is a common way to represent the decision-making process by means of a tree-like and branching structure. A DT is a method for structuring an algorithm. The DT algorithm is used to divide the features of the data set through the cost function. Before optimization and removal of extra branches, this algorithm grows in a way that has features unrelated to the problem. For this reason, "pruning" operation is performed to remove these extra branches. In the DT algorithm, parameters such as the depth of the DT can also be adjusted to prevent overfitting or overcomplication of the tree. More details on DT can be found in (16).

The "Neural Network", which is abbreviated to "NN", is a versatile powerhouse inspired by the human brain's interconnected neurons. It possesses the capacity to tackle a wide spectrum of tasks, including image recognition and natural language processing. As discussed in (17), its deep architecture allows it to model intricate relationships within data, making it a formidable contender in the machine learning landscape.

Table 2. Result of Wilma applied on Hoda data set.

Digits	Accuracy
0 against all	%99
1 against all	%98
2 against all	%96
3 against all	%95
4 against all	%95
5 against all	%99
6 against all	%90
7 against all	%97
8 against all	%99
9 against all	%98

Table 3. Prediction accuracy results of digit labels in Hoda dataset using DT, NN, SVM and Wilma methods. Noted that all the values in the table are multiplied by 100.

	DT	NN	SVM	Wilma
0 against all	71	76	96	**99**
1 against all	95	95	96	**98**
2 against all	90	66	93	**96**
3 against all	76	85	93	**95**
4 against all	57	**95**	93	**95**
5 against all	52	76	**100**	99
6 against all	**95**	85	**95**	90
7 against all	80	85	95	**97**
8 against all	60	90	**100**	99
9 against all	57	85	**98**	**98**

The “Support Vector Machine” with the abbreviation “SVM” is also one of the popular supervised models that was developed by (18) and is used for data classification. The SVM algorithm creates a hyperplane in which the distance between two classes of data points is at its maximum. This hyperplane is known as the decision boundary that separates the classes of data on both sides of the plane.

Each of the supervised algorithms explained above has its own advantages and disadvantages. For example, a DT is a simple model that can be easily explained to those without an analytical background. DT requires less pre-processing for input data than other data modeling methods. At the same time, if the data is not clear and overlaps, the DT technique does not work well. It is also not ideal for large data sets and becomes complex.

In the case of a NN, it stands out for its capability to handle complex relationships within the data. NNs are adept at capturing intricate patterns and non-linear dependencies, making them suitable for tasks with high complexity. However, this increased capability comes at the cost of interpretability, as NNs are often considered “black-box” models. They require more extensive pre-processing of data and can be computationally demanding, particularly with large datasets.

The SVM also does a relatively good time consumption when there is a clear margin of separation between classes. It acts effectively because of its relatively efficient memory in cases where the number of dimensions is greater than the sample. At the same time, when the dataset has a lot of noise, the SVM has a poor rejection.

Now, we examin the supervised procedures described above in order to determine the performance of these algorithms in the Hoda dataset and compare them with the results of the Wilma algorithm. The results are given in Table 3.

In Table 3, we have denoted every method that exhibited higher accuracy for each digit by bold formatting. According to the results in Table 3, all four algorithms have acceptable performance in predicting. Among the four scrutinized algorithms, the Wilma's algorithm consistently exhibits commendable performance. Upon closer examination, it becomes evident that, in the majority of cases, the Wilma's algorithm outperforms the other supervised methods in terms of accuracy. This compelling trend signifies the robust efficacy of the Wilma's algorithm, particularly when confronted with high-dimensional datasets.

Conclusions

The nature and structure of the predictor variables are the only details used by popular clustering methods (as unsupervised tools) to group the data. Better clustering outcomes appear to be achieved when the response variable is included in the clustering job. This article's primary objective was to provide an implementation of supervised clustering in a practical setting. It was demonstrated that cluster analysis produces good results when the relationship between predictor and response variables is taken into account, and the variables' placement inside clusters is not negated by this relationship. When there were more than two response categories, supervised clustering was also looked into. Moreover, it is demonstrated that the accuracy of predicting the observation class label-a crucial topic when working with high-dimensional data-was not compromised by taking into account significant variables and using the dimensionality reduction technique on the Hoda dataset.

In future research, we plan to implement the Pelora algorithm (introduced in (15), which is the L_2 penalized negative log-likelihood function on Hoda dataset. Besides, our another objective is to compare the prediction results by changing the penalty function L_2 to L_1.

Acknowledgments

The authors would like to thank the reviewers and editor for providing constructive comments leading to improve the quality of our manuscript.

References

1. Izenman AJ. Modern multivariate statistical techniques. New York: Springer; 2008.
2. McLachlan GJ, Basford KE. Mixture models: Inference and applications to clustering. New York: M. Dekker; 1988 Jul.
3. Bühlmann P, Van De Geer S. Statistics for high-dimensional data: methods, theory and applications. Springer Science & Business Media; 2011 Jun 8.
4. Fraiman R, Justel A, Svarc M. Selection of variables for cluster analysis and classification rules. Journal of the American Statistical Association. 2008 Sep 1:1294-303.
5. Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. CRC press; 2015 May 7.
6. Dettling M, Bühlmann P. Supervised clustering of genes. Genome biology. 2002 Dec;3:1-5.
7. Hastie T, Tibshirani R, Botstein D, Brown P. Supervised harvesting of expression trees. Genome Biology. 2001 Jan;2(1):1-2.
8. Allwein EL, Schapire RE, Singer Y. Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of machine learning research. 2000;1(Dec):113-41.
9. Liu CL, Nakashima K, Sako H, Fujisawa H. Handwritten digit recognition: benchmarking of state-of-the-art techniques. Pattern recognition. 2003 Oct 1;36(10):2271-85.
10. Mayraz G, Hinton GE. Recognizing hand-written digits using hierarchical products of experts. Advances in neural information processing systems. 2000;13.
11. Trier ØD, Jain AK, Taxt T. Feature extraction methods for character recognition-a survey. Pattern recognition. 1996 Apr 1;29(4):641-62.
12. Suen CY, Nadal C, Legault R, Mai TA, Lam L. Computer recognition of unconstrained handwritten numerals. Proceedings of the IEEE. 1992 Jul;80(7):1162-80.
13. Khosravi H, Kabir E. Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. Pattern recognition letters. 2007 Jul 15;28(10):1133-41.
14. Tibshirani R, Walther G, Hastie T. Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(2):411-23.
15. Dettling M, Bühlmann P. Finding predictive gene groups from microarray data. Journal of Multivariate Analysis. 2004 Jul 1;90(1):106-31.
16. Freund Y, Mason L. The alternating decision tree learning algorithm. Inicml 1999 Jun 27 (Vol. 99, pp. 124-133).
17. Jing L, Tian Y. Self-supervised visual feature learning with deep neural networks: A survey. IEEE transactions on pattern analysis and machine intelligence. 2020 May 4;43(11):4037-58.
18. Cortes C, Vapnik V. Support-vector networks. Machine learning. 1995 Sep;20:273-97.
19. Haponchyk I, Moschitti A. Supervised neural clustering via latent structured output learning: application to question intents. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 2021 Jun (pp. 3364-3374).
20. Yang Y, Teng F, Li T, Wang H, Wang H, Zhang Q. Parallel semi-supervised multi-ant colonies clustering ensemble based on mapreduce methodology. IEEE Transactions on Cloud Computing. 2015 Dec 23;6(3):857-67.