Advances in Computer, Signals and Systems (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2023.070806
ISSN 2371-8838 Vol. 7 Num. 8

WXGCB: A Clustering Prior Weighting Semi-Supervised Learning Method Based on Space Level Constraint and Mixed Variable Metrics

Xinguang Wang[1,a,*]

[1]*School of Information and Electronic Technology, Key Laboratory of Autonomous Intelligence and Information Processing in Heilongjiang Province, Jiamusi University, Jiamusi, China*
[a]*wangxinguang2021@gmail.com*
[*]*Corresponding author*

Keywords: Semi-supervised learning, mixed variable, space-level constraints, clustering prior

Abstract: A clustering prior weighted semi-supervised learning method called WXGCB has been proposed, which combines the characteristics of the cluster-then-label semi-supervised method and space-level constraint semi-supervised method. WXGCB can use mixed variable information, data prior information, and clustering prior information based on different clustering algorithms to adjust the distance matrix, thereby transforming different supervised learning algorithms into semi-supervised learning algorithms for improving their prediction accuracy. Due to the fact that WXGCB does not require internal adjustments to the clustering algorithms and supervised learning algorithms used, this method can flexibly combine different clustering algorithms and supervised learning algorithms to find combinations that can better compensate for each other's shortcomings, and can easily convert various supervised learning algorithms into semi-supervised learning algorithms. To verify the effectiveness of WXGCB, WXGCB transformed two supervised learning algorithms KSNN and DBGLM into semi-supervised mixed variable learning algorithms SMKSNN and SMGLM, and conducted performance comparison experiments with the other two semi-supervised learning algorithms on six benchmark datasets.

1. Introduction

Machine learning addresses the question of how to build computers that improve automatically through experience, and is one of today's most rapidly growing technical fields, lying at the intersection of computer science and statistics, and at the core of artificial intelligence and data science [1]. After about 80 years of development, machine learning has produced numerous branches and achievements [2,3]. Machine learning algorithms can be roughly divided into supervised learning, semi-supervised learning, and unsupervised learning based on the degree of dependence of algorithms on prior information such as data labels in the dataset being studied.

Supervised learning algorithms require sufficient prior information on the data to be learned, which usually means that the data labels of the training set are known. KNN is a classic supervised

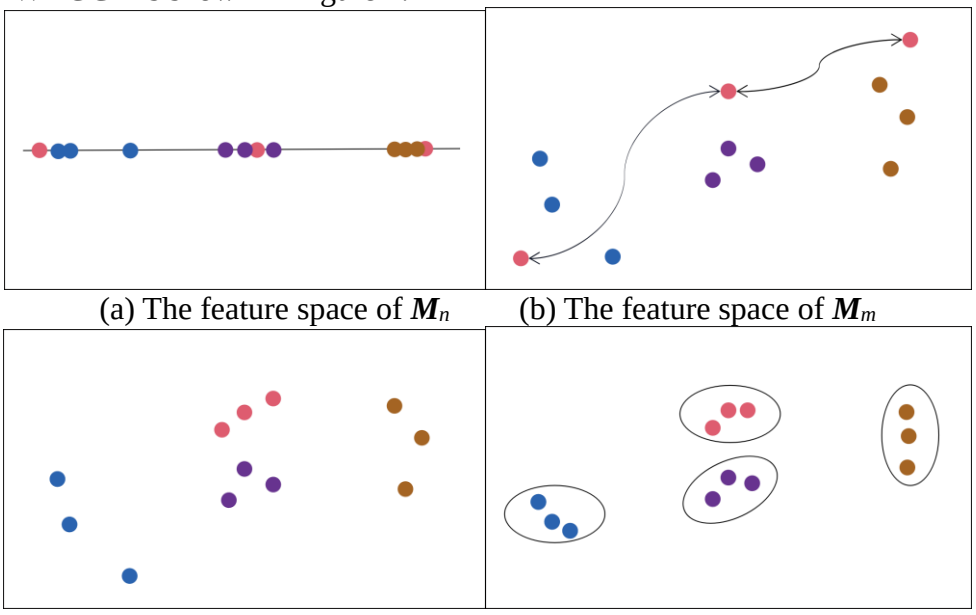
classification algorithm that finds k nearest training vectors $\boldsymbol{x}_i, \boldsymbol{x}_{i+1},...,\boldsymbol{x}_{i+k}$ of the predicted vector $\boldsymbol{x}_j$ and determines that $\boldsymbol{x}_j$ should belong to the same class as the training vector with the highest number of classes among these k vectors [4-6]. KSNN is a powerful variant of KNN, which proposes an effective method to solve the problem of optimal number of neighbors and the optimal weight value [7]. GLM is a classic regression algorithm that proposes a more generalized linear model framework that allows response variables to be not only continuous variables of normal distribution but also other types of distributions, such as binomial distribution, Poisson distribution, etc [8]. DBGLM is a variant of GLM where explanatory information is coded as distances [9]. Unlike GLM, which uses training vector sets or sample feature matrices as inputs, DBGLM can use dissimilarity matrices or distance matrices as inputs and can more naturally use mixed variables. Although these supervised algorithms have superior performance, they are limited by being able to train only on labeled data with sufficient prior information, and cannot further improve model performance by utilizing unlabeled data without prior information.

Unsupervised learning can only utilize the feature vectors of samples and cannot utilize prior information, and clustering algorithms are one of the important branches of unsupervised learning [10]. DBSCAN is a classic density-based clustering algorithm that can effectively handle clusters of different shapes, including non-convex clusters, and identify outlier noise points [11]. HDBSCAN, which is a variant of DBSCAN, can handle clusters of different shapes and densities well, and can adaptively select suitable epsilon neighborhood [12]. Similar to supervised learning, these powerful unsupervised learning algorithms cannot further improve clustering performance by utilizing prior information. Additionally, they are limited by the characteristics of measurement functions and can only utilize a single numerical variable.

Semi-supervised learning algorithms can comprehensively utilize data with prior information and data without prior information to improve algorithm performance. Nowadays, the research field of semi-supervised learning has generated many different branches [13]. Cluster-then-label semi-supervised learning method uses the clustering results generated on all datasets to assist the subsequent supervised learning algorithm, thus comprehensively utilizing data with prior information and data without prior information in the overall algorithm. The graph-based semi-supervised learning method represents data points as nodes in the graph and then constructs edges of the graph-based on their similarity. Then, a prediction model is constructed based on the idea that unlabeled data points adjacent to or similar to labeled data points may belong to the same category. The space-level constraint semi-supervised learning method is similar to the graph based semi-supervised learning method, which achieves soft constraints by transforming prior information into bringing the distance between two data points closer or farther apart [14]. In the field of semi-supervised learning, many achievements have been made, including many powerful semi-supervised learning algorithms [15]. WSVM is a variant of SVM that proposes a new label generation strategy to handle the mixed integer programming problem, thereby improving the model's performance on weakly labeled data [16,17]. MCPL proposes a general way to perform semi-supervised parameter estimation for likelihood-based classifiers on the full training set where the estimates are never worse than the supervised solution in terms of the log-likelihood, and MCPL is applied to LDA to generate an algorithm instance MCPLDA to verify its effectiveness [18,19].

In this paper, a clustering prior weighted semi-supervised learning method called WXGCB has been proposed, which combines the characteristics of the cluster-then-label semi-supervised method and space-level constraint semi-supervised method. Firstly, WXGCB utilizes the information in mixed variables by using mixed variable metric functions to calculate the mixed variable distance matrix $\boldsymbol{M}_m$. Compared to the distance matrix $\boldsymbol{M}_n$ of a single variable, $\boldsymbol{M}_m$ has a higher dimension, resulting in a sparser feature space and clearer boundaries between data points of different categories. Then WXGCB converts the prior information of the training set, such as data labels, into

space-level constraints and weights the corresponding elements in $\boldsymbol{M}_m$ according to these constraints, obtaining a constrained weighted matrix $\boldsymbol{M}_s$ that carries both mixed variable information and data prior information. $\boldsymbol{M}_s$ will be directly inputted into an unsupervised clustering algorithm by WXGCB and the clustering results will be obtained. Due to the prior information and mixed variable information contained in $\boldsymbol{M}_s$, the clustering results may be more reliable. The clustering results will be assumed by WXGCB as weak prior information with lower confidence than the prior information in the dataset and transformed into space-level constraints to further adjust $\boldsymbol{M}_s$, obtaining the clustering prior weighting matrix $\boldsymbol{M}_c$. $\boldsymbol{M}_c$ will be directly used as input to a supervised learning algorithm by WXGCB, which includes weak prior information of the clustering algorithm, prior information of the dataset, and mixed variable information to help this supervised algorithm. For example, since HDBSCAN can effectively cluster clusters of any shape and density and identify outlier noise points, the weak prior information in HDBSCAN-based $\boldsymbol{M}_c$ contains this information, which may supplement the shortcomings of subsequent supervised algorithm KNN. Specifically, it can alleviate the sensitivity of the KNN algorithm to noise outliers and class imbalance and may improve the final prediction results. In addition, due to the mixed variable information in $\boldsymbol{M}_c$, the prediction results for mixed variable datasets may also be improved. The main idea of WXGCB is shown in Figure 1.



(a) The feature space of $\boldsymbol{M}_n$ (b) The feature space of $\boldsymbol{M}_m$

(c) The feature space of $\boldsymbol{M}_s$ (d) The feature space of $\boldsymbol{M}_c$

Figure 1: The main idea of WXGCB.

2. Proposed Method

2.1. Mixed Variable Distance Matrix Based on Mixed Variable Metric Function

First of all, the meaning of mixed variables in this paper will be clearly defined and carefully explained. Stevens measurement scale typology is composed of four types of variables including nominal, ordinal, interval, and ratio according to their statistical meaning and permissible statistics [20]. Nominal variable, which is a unique identifier in the form of colors, letters, etc., is only allowed to calculate mode, numbers, etc. It is not allowed to compare, add, subtract, multiply, and divide. Categorical and binary variables belong to nominal variables. Ordinal variable, which is the determination of sort such as larger and smaller, is further allowed to calculate median, quartile, etc. Ordinal variable value is not allowed to add, subtract, multiply, and divide. The interval variable,

which has an equidistant numerical level between adjacent values based on the ordinal variable, is also allowed to calculate sum, difference, etc. But interval variable value is not allowed to multiply and divide because it has no exact zero such as Celsius degree. A ratio variable with a definite zero is allowed to do any calculation. The numerical variable belongs to the ratio variable. Therefore, the amount of information contained and the types of operation methods are sequentially increasing among the four variable types.

If machine learning algorithms can utilize mixed variables instead of single variables, the accuracy of predictions may improve due to the increase in information. For the distance calculation of mixed variable data, if variables of low-level type are included in the operation of high-level without additional prior information, it is impermissible and meaningless according to Stevens typology. Conversely, including variables of high-level type in the operation of low-level causes information loss. So it is a challenge to measure the distance between mixed data without variable type conversion. To solve this problem, many mixed variable metric functions have been proposed [21]. In this paper, the classic Gower distance is used as an example [22]. The metric function for calculating mixed variables using Gower distance is given by

$$d_{ij} = 1 - s_{ij} = 1 - \frac{\sum_{l=1}^{p} w_{ijl} s_{ijl}}{\sum_{l=1}^{p} w_{ijl}} \tag{1}$$

Where s_{ijl} is the general similarity coefficient between two data points x_i and x_j on feature l, representing the degree of similarity. s_{ij} is the weighted average general similarity coefficient, and in this paper, the weight value defaults to 1. The calculation method for the general similarity coefficient s_{ijl} is as follows

$$s_{ijl} = \begin{cases} 1 - \dfrac{\left| x_{il} - x_{jl} \right|}{x_l^{\max} - x_l^{\min}}, l \in l_r \\ 1, x_{il} = x_{jl} \text{ and } l \in l_n \\ 0, x_{il} \neq x_{jl} \text{ and } l \in l_n \end{cases} \tag{2}$$

Where l_r means ratio variable such as numerical variable while l_n means nominal variable such as binary variable and category variable. Let training set $\boldsymbol{M}_x = (\boldsymbol{x_1},\boldsymbol{x_2},...,\boldsymbol{x_n})^{\mathrm{T}}$ is an object-feature matrix with n rows and p columns, where each object vector $\boldsymbol{x}_i = (x_{i1},x_{i2},...,x_{ip})^{\mathrm{T}}$ is composed of p feature elements. By using mixed variable metric functions such as Gower distance, WXGCB can calculate the mixed variable distance matrix $\boldsymbol{M}_m$ from $\boldsymbol{M}_x$ as follows

$$\boldsymbol{M}_m = \begin{Bmatrix} d_{11} & ... & d_{1n} \\ ... & ... & ... \\ d_{n1} & ... & d_{nn} \end{Bmatrix} = \begin{Bmatrix} 0 & ... & d_{1n} \\ ... & ... & ... \\ d_{n1} & ... & 0 \end{Bmatrix} \tag{3}$$

Where 0 indicates that the distance from the object itself to itself should be 0. Compared to the single variable distance matrix $\boldsymbol{M}_n$ calculated from the features of a single variable, the algorithm using $\boldsymbol{M}_m$ is more likely to have higher prediction accuracy. On the one hand, this may be because $\boldsymbol{M}_m$ contains more relevant information, and on the other hand, it may be because the mixed variable feature space has a higher dimension than the single variable feature space, resulting in a

more sparse feature space where different data points in different data clusters are easier to distinguish from each other.

2.2. Constrained Weighted Distance Matrix Based on Space Level Constraint

In semi-supervised learning methods, prior information refers to existing prior knowledge or additional information about data, usually obtained from the context of domain knowledge, domain experts, other data sources, or problems. The types of prior information usually include data labels, similarity relationships between data points, data distribution, and so on. In terms of the representation of prior information, space-level constraint method are similar to instance-level constraint method, which represent prior information as paired constraints. Paired constraints can indicate whether two objects belong to the same or different categories. Specifically, if two objects belong to the same category, a must link (ML) constraint can be generated between them; On the contrary, if two objects belong to different categories, a cannot link (CL) constraint can be generated between them. However, the space-level constraint method further transforms pairing constraints into correcting the distance between data points in the feature space, thereby achieving soft constraints and alleviating constraint violation issues. Specifically, the distance between two data points with ML constraint will be reduced, while the distance between two data points with CL constraint will be enlarged. To utilize this idea, WXGCB converts CL and ML constraints into weights and weights the elements in the $\boldsymbol{M}_m$ matrix to change the distance between relevant data points, thereby representing prior information as an adjustment to the distance matrix. In this way, the constrained weighted distance matrix $\boldsymbol{M}_s$ containing prior information and mixed variable information can be given by

$$\boldsymbol{M}_s = \boldsymbol{M}_m \circ \boldsymbol{M}_{w^s} = \begin{Bmatrix} 0 & \cdots & d_{1n} w_{1n}^s \\ \cdots & \cdots & \cdots \\ d_{n1} w_{n1}^s & \cdots & 0 \end{Bmatrix} \tag{4}$$

Where w^s is $\boldsymbol{M}_s$ is the weight obtained by converting two paired constraints, while $\boldsymbol{M}_{w^s}$ is a weight matrix composed of all weights. w^s is given by

$$w_{ij}^s = \begin{cases} v^s, ij \in CL \\ 1/v^s, ij \in ML \\ 1, \text{otherwise} \end{cases} \tag{5}$$

Where v^s is a fixed weight value greater than 1, used to represent the strength of the constraint, which defaults to 4 in this paper.

2.3. Clustering Prior Weighting Matrix Based on Clustering Results

Cluster-then-label semi-supervised learning method is a common strategy in semi-supervised learning. Its main idea is to first cluster unlabeled data, and then assign labels to the data points in the cluster based on the clustering results. The goal of this method is to enhance the performance of the model by utilizing the inherent distribution structure of the data. Similarly, WXGCB also uses clustering results to correct the training data of subsequent supervised learning algorithms, making the dataset more likely to conform to the internal distribution structure, thereby improving model performance. Due to being directly used as input to the clustering algorithm by WXGCB, the mixed variable information and prior information contained in $\boldsymbol{M}_s$ may make the clustering results more

reflective of the potential distribution structure of the data. In WXGCB, clustering results are considered weak prior information. Similar to the utilization of prior information, weak prior information will be transformed into ML and CL constraints, and further transformed into weights as follows

$$w_{ij}^{c}=\begin{cases}v^{c}, ij\in CL\\ 1/v^{c}, ij\in ML\\ 1, \text{otherwise}\end{cases} \tag{6}$$

Where v^c is a fixed weight value greater than 1, used to represent the strength of the constraint, which defaults to 2 in this paper. Considering that the clustering results belong to weak prior information, their confidence level is usually lower than the prior information of the dataset. Therefore, a weight value smaller than v^s is usually a suitable choice. Then, similar to the calculation of $\boldsymbol{M}_s$, the clustering prior weight matrix $\boldsymbol{M}_c$ is the Hadamard product of $\boldsymbol{M}_s$, and the clustering prior weight matrix $\boldsymbol{M}_w{}^c$ is as follows

$$\boldsymbol{M}_c=\boldsymbol{M}_s\circ\boldsymbol{M}_{w^c}=\begin{Bmatrix}0 & \cdots & d_{1n}w_{1n}^{c}\\ \cdots & \cdots & \cdots\\ d_{n1}w_{n1}^{c} & \cdots & 0\end{Bmatrix} \tag{7}$$

Since $\boldsymbol{M}_c$ is weighted on the basis of $\boldsymbol{M}_s$, it not only includes clustering prior information, but also preserves mixed variable information and prior information of the dataset. In fact, prior clustering information may reinforce the additional information.

WXGCB directly uses $\boldsymbol{M}_c$ as an input to a distance-based supervised learning algorithm to transform it into a semi-supervised learning algorithm that can utilize mixed variables, dataset prior information, and clustering prior information. Due to the lack of internal changes to this clustering algorithm and this supervised learning algorithm during this process, different supervised learning algorithms can be directly transformed from WXGCB to semi-supervised learning algorithms, and WXGCB can easily use different combinations of clustering algorithms and supervised learning algorithms to adapt to different datasets and learning objectives. In this paper, we present two combination examples in WXGCB, one using a combination of HDBSCAN and KSNN, and the other using a combination of HDBSCAN and DBGLM. HDBSCAN can effectively cluster clusters of different shapes and densities and identify outlier noise points. Therefore, the clustering of prior information generated by HDBSCAN may effectively alleviate some of the problems of supervised learning algorithms, such as outlier sensitivity and class imbalance sensitivity.

3. Experiment and Result

3.1. Experimental Design

3.1.1. Comparison Algorithms

In this paper, WXGCB selects Gower distance as the mixed variable metric function and HDBSCAN as the clustering algorithm for generating weak prior information as a version of WXGCB. Then WXGCB converts KSNN into SMKSNN, and DBGLM into SMGLM as two examples of semi-supervised learning algorithms transformed by WXGCB. In order to verify the effectiveness of WXGCB, SMKSNN and SMGLM will conduct comparative experiments with their respective original algorithms SKNN and GLM on the benchmark datasets, and further

conduct comparative experiments with semi-supervised learning algorithms WSVM and MCPLDA on the same benchmark datasets. In addition, the prototype algorithm KNN of KSNN and the prototype algorithm SVM of WSVM are also included in comparative experiments. In the experiment, KNN is implemented by R package class, KSNN is implemented by R package KsNN, DBGLM is implemented by R package dbstats, SVM is implemented by R package e1071, WSVM and MCPLDA are implemented by R package RSSL, HDBSCAN is implemented by R package dbscan, and Gower distance is implemented by R package kmed. The entire experiment was run in R version 4.1.3 with an Intel Core i7-10870H CPU and 16GB RAM.

3.1.2. Benchmark Datasets

The SMKSNN and SMGLM semi-supervised learning algorithms obtained from WXGCB transformation will be compared with the six comparison algorithms mentioned above on benchmark datasets as Table 1 shows. The benchmark dataset includes four artificial datasets with different characteristics as Figure 2 shows and two real-world datasets from UC Irvine Machine Learning Repository. The true label values of all datasets are set as a sequence of integer values that gradually increase from 1 to the number of clusters.

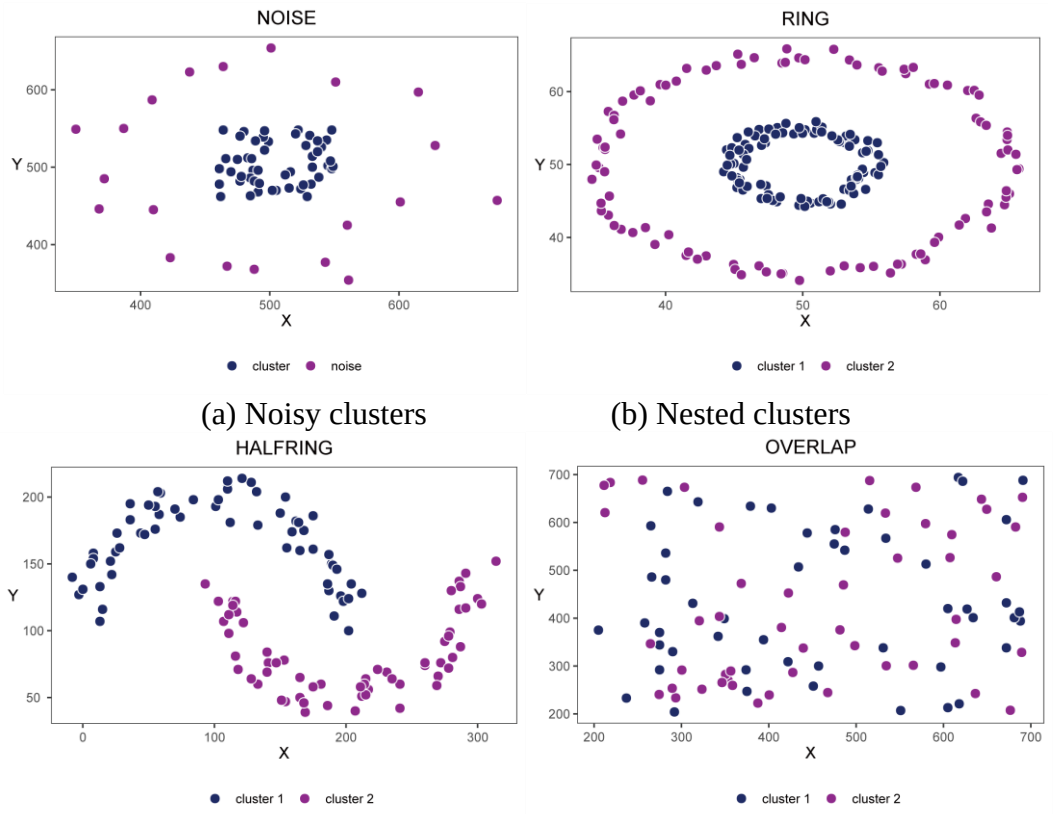
Table 1: Benchmark Datasets.

Dataset	n	k	p_n	p_b	p_c
NOISE	70	2	2	0	0
RING	150	2	2	0	0
HALFRING	120	2	2	0	0
OVERLAP	100	2	2	1	0
FLAG	194	8	10	12	6
FERTILITY	100	2	2	3	4

Where n is object quantity, k is cluster quantity, p_n is numerical variable dimension, p_b is binary variable dimension, p_c is categorical variable dimension.

FLAG is a mixed variable dataset with real-world data about various nations around the world and their corresponding flags. This dataset consists of 194 objects and their 30-dimensional mixed variable features. In this experiment, the predicted feature is set to the religious type of the country to which the flag belongs. Due to the inclusion of 8 religious types in this feature, the actual number of clusters in this dataset is 8. The number of objects contained in each category ranges from 4 to 60, so FLAG is essentially an imbalanced dataset in terms of categories. In addition, the feature about the country name to which the flag belongs is deleted because each object has its own unique country name, making this feature meaningless for predicting the target. The remaining 10 numerical variables, 12 binary variables, and 6 category variables are used for algorithm learning.

FERTILITY is also a mixed variable dataset about real-world medical examination results with 100 objects on 9-dimensional mixed variables and is divided into two categories by medical experts. One class contains 12 objects while the other class contains 88 objects, so this is also an imbalanced dataset in terms of categories. The features used for algorithm learning include 4 category variables, 3 binary variables, and 2 numerical variables.



(a) Noisy clusters (b) Nested clusters

(c) Non-convex clusters (d) Overlapping clusters

Figure 2: Artificial datasets.

NOISE is an artificial dataset containing two-dimensional numerical variable features. It can be divided into two clusters, namely a compact real cluster with 50 objects and a scattered noise cluster with 20 objects as background noise (Figure 2a). Therefore, essentially NOISE is a noisy dataset.

RING is an artificial dataset containing two-dimensional numerical variable features, consisting of two concentric circles, with a small circle containing 50 objects and a large circle containing 100 objects (Figure 2b). Therefore, RING is essentially a class imbalanced dataset.

HALFRING is also an artificial dataset containing two-dimensional numerical variable features, consisting of two semi-circular clusters with similar sizes that are relative to each other and interlaced with each other without any overlap (Figure 2c). Therefore, HALFRING is a non-convex cluster dataset with two clusters each cluster containing 60 objects.

OVERLAP is a mixed variable artificial dataset consisting of two numerical variables and one binary variable. In a feature space composed of two numerical variables, it is composed of two overlapping clusters (Figure 2d). However, the boundary between the two clusters on the binary variable is very clear, so OVERLAP is a dataset of overlapping clusters but with distinguishable mixed variable information.

3.1.3. Parameters Settings

Each benchmark dataset is divided into two halves, with half being used as the training set to train the models and the other half as the testing set to verify the prediction accuracy of the models. Then, the training set uses 5-fold cross-validation to obtain the appropriate hyperparameters for each algorithm. Before dividing the dataset and cross-validation, in order to ensure that each subset can retain the inherent data structure of the original benchmark dataset, the experiment randomly

divided the original benchmark dataset into 10 folds using stratified sampling. For cases where the number of objects in a certain layer of a fold is not an integer multiple of 10, the experiment uses random oversampling to supplement the number of objects.

In terms of parameter optimization of the algorithm. MCPLDA does not use 5-fold cross-validation but trains the model directly on the training set according to the default maximum number of iterations parameter. DBGLM also directly trains the model on the training set, using Poisson regression models and logarithmic linking functions to establish response variables and linear predictions, and using effective rank to fit the generalized linear model. SMGLM is the same as DBGLM in setting these parameters but requires the 5-fold cross-validation to select the hyperparameter of HDBSCAN, which is the minimum number of neighbors *MinPts*. The tuning range is from 2 to the number of objects. Similarly, SMKSNN also needs to find the optimal *MinPts* within this range through cross-validation. In addition, SMKSNN also needs to find the optimal number of neighbors *k* by 5-fold cross-validation, ranging from 1 to the number of objects. KSNN and KNN also use the same range to find suitable *k* values. SVM and WSVM use linear kernels and use 5-fold cross-validation to find the optimal cost of constraints violation, with a search range of 2^{-8}, 2^{-7},..., 2^{0},..., 2^{7}, 2^{8}.

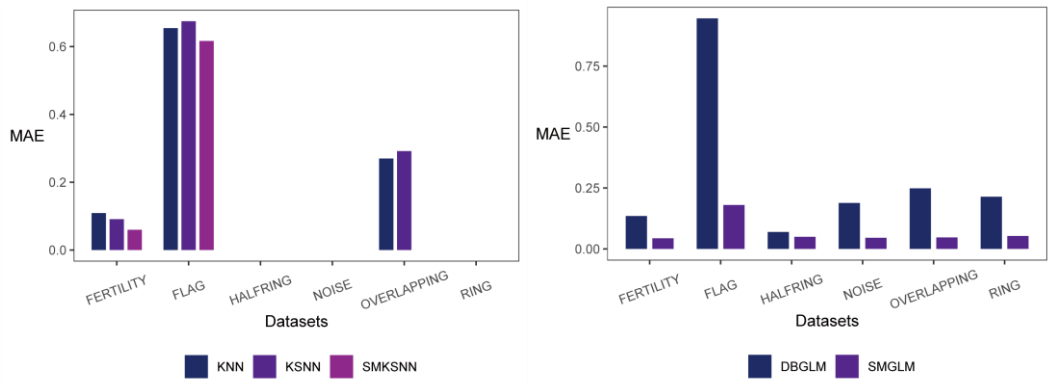
3.1.4. Parameters Settings

Due to the inclusion of both regression and classification algorithms in the experiment, the predicted results are either continuous variable values or integer values. In order to facilitate and accurately compare these results, the evaluation index of the experimental results is selected as the mean absolute error (MAE) as follows

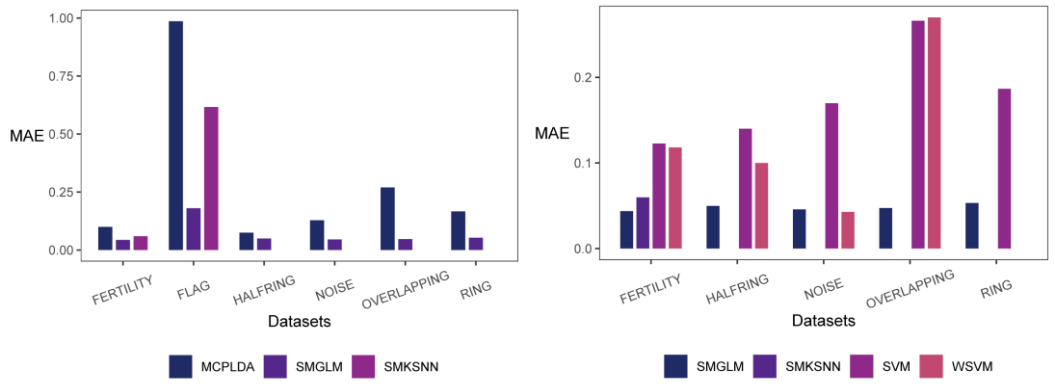
$$MAE = \frac{\sum_{i=1}^{p} \left| y_i - y_i^r \right|}{p} \tag{8}$$

Where y_i is the real label of object i while y_i^r is the predicted label of the same object. MAE can represent the average degree to which all predicted results of a prediction model deviate from the actual results.

3.2. Experimental Result



(a) KNN, KSNN, and SMKSNN (b) DBGLM and SMGLM



(c) MCPLDA, SMKSNN, and SMGLM (d) SVM, WSVM, SMKSNN, and SMGLM

Figure 3: Experimental result.

The experimental result of the comparison of SMKSNN with its prototype algorithms KSNN and KNN is shown in Figure 3a. SMKSNN has the lowest MAE on all datasets, which means that SMKSNN has the lowest degree of deviation from prior results in terms of prediction accuracy.

The experimental result of the comparison of SMGLM with its prototype algorithms DBGLM is shown in Figure 3b. In this experiment, SMGLM has lower MAE than DBGLM in all datasets, which means that SMGLM has the lowest degree of deviation from prior results in terms of prediction accuracy.

The comparison results of SMKSNN, SMGLM, and semi-supervised learning algorithm MCPLDA are shown in Figure 3c. The MAE of SMGLM and SMKSNN is lower than that of MCPLDA in all datasets. In addition, SMKSNN performs better than SMGLM on datasets HALFRING, NOISE, OVERLAP, and RING, while the opposite is true on datasets FLAG and FERTILITY.

The comparison results of SMKSNN, SMGLM, SVM, and semi-supervised learning algorithm WSVM are shown in Figure 3d. Due to the fact that WSVM can only directly handle binary classification tasks, these experimental results are limited to binary classification datasets. On all datasets, the results of SMKSNN and SMGLM outperform those of SVM and WSVM.

4. Conclusions

The following conclusions can be drawn from the experimental results. Firstly, SMKSNN and SMGLM demonstrate better predictive accuracy than their respective prototype algorithms KSNN and DBGLM in these benchmark datasets. This indicates that the prediction accuracy of supervised learning algorithms that can utilize mixed variable information, data prior information, and clustering prior information may be improved by transforming WXGCB into a semi-supervised clustering algorithm. In addition, the predictive accuracy of SMKSNN and SMGLM in these benchmark datasets is attributed to the semi-supervised learning algorithms MCPLDA and WSVM, which suggests that the semi-supervised learning algorithm transformed by WXGCB may have better predictive accuracy than traditional and specialized semi-supervised learning algorithms. The WXGCB in this paper is only an example of WXGCB. By improving the Gower distance to other potential metric functions that are more suitable for certain mixed variable datasets, and selecting a combination of other clustering algorithms and supervised learning algorithms, the performance of WXGCB may be further improved on certain datasets. WXGCB, a distance matrix based method, also provides a new feasible research direction for quickly transforming supervised learning algorithms into semi-supervised learning algorithms to utilize additional information to improve the performance of the original algorithm.

References

[1] Jordan M I, Mitchell T M. (2015) Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[2] Mahesh B. (2020) Machine learning algorithms-a review. International Journal of Science and Research (IJSR). [Internet], 9(1), 381-386.

[3] Fradkov A L. (2020) Early history of machine learning. IFAC-PapersOnLine, 53(2), 1385-1390.

[4] Cover T, Hart P. (1967) Nearest neighbor pattern classification. IEEE transactions on information theory, 13(1), 21-27.

[5] Fix E, Hodges J L. (1989) Discriminatory analysis. Nonparametric discrimination: Consistency properties. International Statistical Review/Revue Internationale de Statistique, 57(3), 238-247.

[6] Wu X, Kumar V, Ross Quinlan J, et al. (2008) Top 10 algorithms in data mining. Knowledge information systems, 14(1), 1-37.

[7] Anava O, Levy K. (2016) k-nearest neighbors: From global to local. Advances in neural information processing systems, 29.*

[8] Nelder J A, Wedderburn R W. (1972) Generalized linear models. Journal of the Royal Statistical Society Series A: Statistics in Society, 135(3), 370-384.

[9] Boj E, Delicado P, Fortiana J. (2010) Distance-based local linear regression for functional predictors. Computational Statistics Data Analysis, 54(2), 429-437.

[10] Ghosal A, Nandy A, Das A K, et al. (2020) A short review on different clustering techniques and their applications. Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018, 69-83.

[11] Ester M, Kriegel H-P, Sander J, et al. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. kdd, 96(34), 226-231.

[12] Campello R J, Moulavi D, Sander J. (2013) Density-based clustering based on hierarchical density estimates. Pacific-Asia conference on knowledge discovery and data mining, 160-172.

[13] Van Engelen J E, Hoos H H. (2020) A survey on semi-supervised learning. Machine learning, 109(2), 373-440.

[14] Klein D, Kamvar S D, Manning C D. (2002) From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. ICML, 2, 307-314.

[15] Reddy Y, Viswanath P, Reddy B E. (2018) Semi-supervised learning: A brief review. Int. J. Eng. Technol, 7(1.8), 81.

[16] Li Y-F, Tsang I W, Kwok J T, et al. (2013) Convex and scalable weakly labeled SVMs. Journal of Machine Learning Research, 14(7).

[17] Cortes C, Vapnik V. (1995) Support-vector networks. Machine learning, 20(3), 273-297.

[18] Loog M. (2015) Contrastive pessimistic likelihood estimation for semi-supervised classification. IEEE transactions on pattern analysis and machine intelligence, 38(3), 462-475.

[19] Blei D M, Ng A Y, Jordan M I. (2003) Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[20] Stevens S S. (1946) On the theory of scales of measurement. Science, 103(2684), 677-680.

[21] Bishnoi S, Hooda B. (2020) A survey of distance measures for mixed variables. International Journal of Chemical Studies.

[22] Gower J C. (1971) A general coefficient of similarity and some of its properties. Biometrics, 857-871.