Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

KNN-SVM Classifiers in Complex Diagnosis

To cite this article: Hao Cheng 2024 *J. Phys.: Conf. Ser.* **2694** 012081

View the article online for updates and enhancements.

You may also like

- Classification study for the imbalanced data based on Biased-SVM and the modified over-sampling algorithm
Liumei Zhang, Baoyu Tan, Tianshi Liu et al.
- Phonocardiography-based mitral valve prolapse detection with using fractional fourier transform
Mahtab Mehrabbeik, Saeid Rashidi, Ali Fallah et al.
- Supervised machine learning tools: a tutorial for clinicians
Lucas Lo Vercio, Kimberly Amador, Jordan J Bannister et al.

This content was downloaded from IP address 5.74.14.247 on 02/09/2024 at 12:58

KNN-SVM Classifiers in Complex Diagnosis

Hao Cheng

National Academy of Innovation Strategy, China Association for Science and Technology, Fuxing Road 3, Haidian District, Beijing, China
Email: chenghao0524@yeah.net

Abstract. In many applications, classification plays an indispensable role due to its powerful detection and diagnosis function. Especially in real data on disease, the detection of important factors and the diagnosis of the result usually bring huge contributions to patients. Simultaneously, complex problems in real data such as imbalanced data and missing data also lead to more challenges and difficulties. The ignorance of missing data will undermine study efficiency, and sometimes introduce substantial bias. Imbalanced data tends to be overwhelmed by the majority classes and ignoring the minority ones. The paper develops new support vector machine classifiers using k-nearest neighbors' information (KNN-SVM), to impute missing data by calculating k-nearest neighbors' statistical characteristic values and to interpolate some new samples between k-nearest minority class examples. As comparisons, the paper uses different kernel functions in KNN-SVM classifiers to show the different performances in disease diagnosis accuracy.
Keywords. Classification, imbalanced data, missing data, kernel.

1. Introduction

Classification belongs to one of the widely discussed topics in many fields. From the perspective of detection and diagnosis, lots of classifiers have been well-developed under the assumptions that the training sets are well-balanced, completely-observed and all misclassification errors cost equally. Well-balanced data means all the classes have relatively equal samples rather than some classes have much more samples than the others. Completely-observed data means each variable and each sample is fully observed without missing data. However, in real data world, data is usually imbalanced and partly missing, causing challenges and difficulties to standard classifiers. More specifically, imbalanced data problem tends to be overwhelmed by the majority classes and ignoring the minority ones, and missing data problem may lead to less efficiency and biased findings which do not represent all the subjects.

In imbalanced data sets, the classes having more examples are defined as the majority classes and the ones having fewer examples as the minority classes. In various real-world settings, the class imbalance problem often occurs in classification investigation due to different reasons [1]. In medical record databases such as whether the patients are dead or not after operation, a large number of patients will belong to the classification "live" rather than "die", which is often treated as the minority class. In practical applications, the ratio of the minority classes to the majority classes can vary from 0 to 1, drastically. Here 0 represents that minority classes do not have any sample at all, and 1 represents that minority classes have the same number of samples as the majority classes.

The basic idea of dealing with imbalanced data is to generate new minority class samples based on k-nearest neighbours' information, and thus obtain relatively equal numbers of samples in both majority and minority classes. There already exist several methods to accomplish the balance through

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

under-sampling the majority class, over-sampling minority class, or both [2-4]. Considering the information loss caused by under-sampling methods, the paper considers the existing synthetic minority over-sampling technique (SMOTE). Its core part is generate new samples belonging to k-nearest neighbours' minority classes through interpolating between the existing minority class samples which lie together [5].

Another common-seen data problem that we mentioned above is missing data. Ignoring the missing data will undermine study efficiency, and sometimes introduce substantial bias. In the case of categorical variables, relatively limited methods can be used to impute new values. In this paper, we consider k-nearest neighbours imputation method (KNN) based on the most frequent value due to the case of categorical variables in real data [6].

As a widely used tool for classification, the support vector machine (SVM) was firstly motivated by the geometric consideration of maximizing the margin [7]. The basic principle of SVM is to find a hyperplane that separates the two classes of data points. There also exist many other classifiers in data mining and machine learning domains such as random forest, neural networks and lasso [8]. In our paper, we focus on classical SVM classifier as basic model and develop two kinds of new support vector machine classifiers using k-nearest neighbours' information (KNN-SVM), which simultaneously deal with imbalanced data and missing data in classification investigations.

More specially, the paper develops the first kind of KNN-SVM classifiers to impute missing data by calculating k-nearest neighbours' means or modes and to interpolate some new samples between k-nearest minority class examples, and then to classify the patients in two groups according to relatively balanced and completely observed data, which belongs to the main contribution of our paper [9]. As comparisons, the paper uses different kernel functions in KNN-SVM classifiers to show the different performances in disease diagnosis accuracy. It should be noted that, the paper does not develop the KNN-SVM diagnosis tools and methods in both static and dynamic settings [10, 11].

The remainder of the paper is organized as follows. We describe our theoretical investigation in Section 2, including the response variable, covariates and settings in Section 2.1, and missingness in Section 2.2, respectively. In Section 3, we introduce the KNN-SVM classifiers. Then Section 4 applies our KNN-SVM classifiers to real data analysis and compared their performances in disease diagnosis accuracy. Some final discussions are placed in Section 5. Figure 1 presents our investigation flowchart.

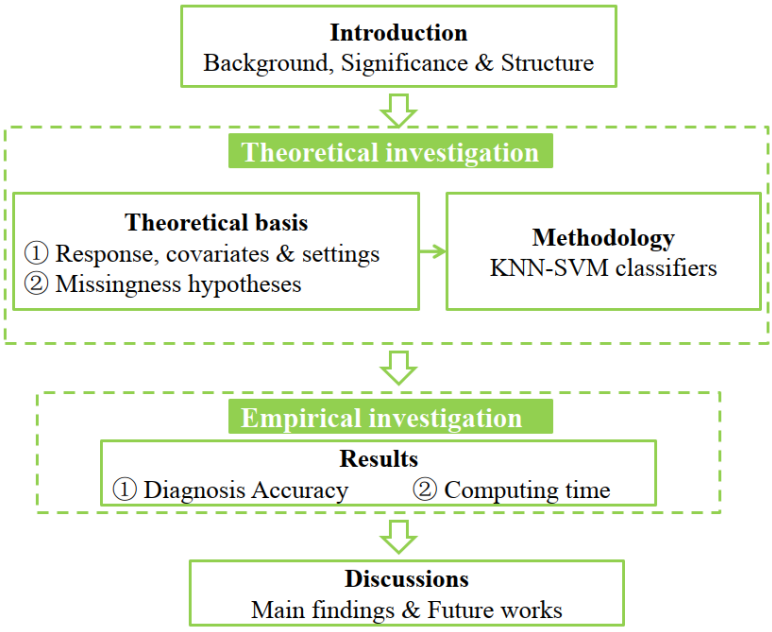

Figure 1. Investigation flowchart.

2. Theoretical Basis Research

2.1. Response, Covariates and Settings

The real data is part of a surgery data from one medical institute, containing 552 patients and 38 variables. The response variable, which is a binary variable (equals 1 or 0), represents whether the patients are currently dead or not. 1 represents the minority class "die" with 101 samples, and 0 represents the majority class "live" with 451 samples. It is easily to calculate the imbalance rate as 451/101 (4.47:1). The covariates represent 36 physical and chemical indicators from 552 patients. For brevity, we denote the setting with imbalanced rate 4.47:1 from 552 patients as **Setting One**.

To further investigate the performances of our proposed KNN-SVM, we randomly delete 50 samples which belong to minority class, and thus we get another sub-dataset containing 502 patients. Obviously, the imbalance rate here is 451/51 (8.84:1), and we denote this setting as **Setting Two**.

In both **Setting One** and **Setting Two**, the response Y and 36 covariates are unchanged. All the covariates and their abbreviations can be seen in table 1.

Table 1. All the covariates and their corresponding abbreviations.

Variable	Abbre.	Variable	Abbre.
myocardial infarction history	XA3	valvular heart disease	XA4
heart failure grade	XA5	Lee's grade	XA1
heart failure	XA2	take aspirin	XB2
use warfarin	XB3	use enoxaparin	XB1
lower limb vein ultrasound	BS	XC with diabetes	XD
with diabetic nephropathy	XE	with tumor	XF1
with other internal medicine diseases	BS	XF2 tumor nature	XF3
smoke history	XG	sex	XH
with COPD	XI1	with respiratory failure	XI2
pulmonary infection within 3 months BS	XI3	COPD severity grade	XJ1
respiratory failure type	XJ2	PLT admission	XK1
myocardial infarction duration	XK5	cerebral infarction duration	XK6
cerebral hemorrhage duration	XK7	ALB admission	XK8
hypertension duration	XK2	coronary heart disease duration	XK3
heart failure duration	XK4	diabetes duration	XL
WBC admission	XM	dementia history	XN
age	XO1	HCT admission	XO2
creatinine admission	XP1	urea admission	XP2

2.2. Missingness Hypotheses

Having investigated the statistical characteristics of all 36 covariates, XA1, XE, XK1, XK8, XM, XN, XO2, XP1 and XP2 have observations with missing values. Table 2 displays the missing rates of the above nine covariates and their corresponding types in two settings **Setting One** and **Setting Two**.

As has mentioned in former section, the ignorance of missingness in input features may lead to potential bias or less efficiency. The common-used approach is to impute the missing values in some way. In the paper, we assume that the missingness is missing at random (MAR). The MAR can be characterized by the conditional distribution of the missing data indicator matrix I given x, say $f(I|x,\phi)$. More specifically, the missingness depends only on the observed components and not on the components that are missing [9]. That is,

$$f(I|x,\phi) = f(I|x_{obs},\phi) \tag{1}$$

where x represents all the covariates vector, x_{obs} represents the observed components and x_{mis} represents the missing components. ϕ represents unknown parameter. I represents the missing data indicator matrix, in which the elements equal 1 which missing, otherwise 0.

Table 2. The covariates XA1, XE, XK1, XK8, XM, XN, XO2, XP1, XP2 and their corresponding missing rates (MR, %) in **Setting One** and **Setting Two**.

Settings	Covariates	MR（%）	Covariates	MR（%）	Covariates	MR（%）
One	XA1	14.1	XE	15.6	XK1	11.8
	XK8	13.6	XM	11.8	XN	5.6
	XO2	11.8	XP1	14.1	XP2	14.1
Two	XA1	15.3	XE	14.7	XK1	12.9
	XK8	14.7	XM	12.9	XN	5.2
	XO2	12.9	XP1	15.3	XP2	15.3

3. Methodology

3.1. Notations

Suppose $x_{i1}, \dots, x_{ip} (i = 1, \dots, n)$ denotes input covariates. $x_{i1}, \dots, x_{ip_1}$ are completely observed covariates. $x_{ip_{1+1}}, \dots, x_{ip}$ are covariates containing missing data, $x_{ip_{1+1}}{}^{(new)}, \dots, x_{ip}{}^{(new)}$ are the corresponding covariates that the missing parts have been imputed using k-nearest neighbours' information. $y_i (i = 1, \dots, n)$ denotes the binary output class label (the response). $y_1, \dots, y_{n_1}$ represent the majority class with relatively large numbers of samples, $y_{n_1+1}, \dots, y_n$ represent the minority class with relatively small numbers of samples and $y_{n_1+1}, \dots, y_N$ represent the new samples belonging to minority class. M is the width of the margin, C is the nonnegative tuning parameter that controls the overlap. $\varepsilon_1, \dots, \varepsilon_n, \varepsilon_{n+1}, \dots, \varepsilon_N$ are slack variables and allow individual observations to be on the wrong side of the hyperplane.

3.2. KNN-SVM Classifiers

Support vector machine classifiers using k-nearest neighbours' information (KNN-SVM) can be summarized as an optimization problem in imbalanced and partly missing data cases. The main idea of KNN-SVM consists of three steps:

Step 1 Impute missing data by calculating k-nearest neighbours' characteristic values referring to the existing k-nearest neighbours imputation method. That is, KNN-SVM uses the k-nearest neighbours to fill in the missing data in a data set. For each case with any missing data, it will search for its k most similar cases and use these cases' values to fill in the missing parts. In our paper, we choose k = 5 and 10.

Step 2 Interpolate some new samples between k-nearest minority class examples referring to the existing synthetic minority over-sampling technique (SMOTE). In other words, SMOTE generate new samples through the following way: Calculate the difference between the target feature sample and its nearest neighbour. Multiply this difference by a random number between 0 and 1, and add it to the target feature vector. This leads to the selection of a random point along the line segment between two specific features. The new generated samples increase the samples belonging to the minority class, which help to effectively force the corresponding decision region to become more general. In our paper, we choose k = 5.

Step 3 Run support vector machine procedure on relatively balanced and completely observed data. Therefore, KNN-SVM can be written as the following equations.

$$max_{\beta_0,\dots,\beta_p,\varepsilon_1,\dots,\varepsilon_n,\varepsilon_{n+1},\dots,\varepsilon_N} M$$

$$subject\ to\ \sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p_1} x_{i,p_1} + \beta_{p_{1+1}} x_{i,p_{1+1}}{}^{(new)} + \cdots + \beta_{p_{1+1}} x_{i,p_{1+1}}{}^{(new)}) \geq M(1 - \varepsilon_i)$$

$$\varepsilon_i \geq 0$$

$$\sum_{i=1}^{N} \varepsilon_i \leq C$$

Based on all above equations, the KNN-SVM classifiers enlarge the feature spaces in specific ways using kernels. In our paper, we take linear, polynomial and radial kernels in total. Algorithm 1 displays the details of KNN-SVM classifiers.

Algorithm 1. KNN-SVM classifiers

Step 1:	Impute missing data according to the existing k-nearest neighbours imputation method.
Step 2:	Interpolate some new samples between k-nearest minority class examples.
Step 3:	Run support vector machine procedure on relatively balanced and completely observed data.
	Step 3-1: Choose different kernels: linear kernel, polynomial kernel and radial kernel.
	Step 3-2: Output confusion matrixes and calculate the accuracies.
Step 4:	Compare the accuracies, computing times and choose the most appropriate KNN-SVM classifier.

4. Empirical Investigation

4.1. Accuracy Comparisons

Traditionally, accuracy is one of the commonly used measures and plays a crucial role in both assessing the classification performance and guiding the classifier modelling. Due to the binary categorical response variable, the paper investigates two-class problem. In imbalanced domains, most studies mainly concentrate on two-class problem. This is because that multi-class problem can be simplified to a class of two-class problems. By convention, accuracy measure is calculated based on confusion matrix, which can be seen in table 3.

Table 3. A confusion matrix for a two-class classification

	Predicted as positive	Predicted as negative
Actual positive class	True Positive (TP)	False Negative (FN)
Actual negative class	False Positive (FP)	True Negative (TN)

Table 3 displays a confusion matrix based on the two-class problem. The first column in table 3 is the actual class label of the samples: Actual positive class and actual negative class. The first row in table 3 presents their predicted class label: Predicted as positive and predicted as negative. True Positive (TP) and True Negative (TN) respectively denote the number of positive and negative samples that are classified correctly, while False Negative (FN) and False Positive (FP) denote the number of misclassified positive and negative samples, respectively.

$$Accuracy = (TP + TN)/(TP + FN + FP + TN) \quad (2)$$

Table 4 illustrates the diagnosis accuracies which are calculated by the above equation using KNN-SVM classifiers. As comparisons, the paper directly deletes the samples with missing data before running KNN-SVM classifier. This is similar to completely cases analysis (CC). Therefore, we denote this competitor as CC-SVM.

In table 4, we can conclude the following findings: (1) The classifiers with kernel "radial" always have relatively better performances in accuracy than the classifiers with other two kernels "linear" and

"polynomial". (2) KNN-SVM classifiers with kernel "polynomial" or "radial" perform better than CC-SVM with corresponding kernels. Different from the kernel "linear", both "polynomial" and "radial" are suitable for nonlinear classification problems. It illustrates that nonlinear KNN-SVM classifiers are more powerful than nonlinear CC-SVM due to the advantages of "polynomial" and "radial" kernels. (3) In **Setting One**, KNN-SVM classifiers with K = 10 are a little bit more accurate than the classifiers with K = 5. However, the increase of K does not bring obvious and stable effects on improvement of diagnosis accuracies.

Table 4. Diagnosis accuracies using KNN-SVM classifiers and their competitors in **Setting One** and **Setting Two**.

		Kernels		
Classifiers	K	linear	polynomial	radial
Setting One				
CC-SVM	-	0.807	0.724	0.889
KNN-SVM	5	0.788	0.736	0.891
KNN-SVM	10	0.797	0.742	0.896
Setting Two				
CC-SVM	-	0.821	0.930	0.997
KNN-SVM	5	0.809	0.800	0.938
KNN-SVM	10	0.810	0.778	0.930

CC-SVM, KNN-SVM only using completely observed samples.

4.2. Computing Time

In this section, we investigate the computing time of different KNN-SVM classifiers and their competitors, which can be seen in table 5. As expected, the computing time using KNN-SVM classifiers is three times that of using CC-SVM. However, all classifiers' computing time is less than 1.2 seconds, and with the increase of K from 5 to 10, the KNN-SVM classifiers are comparable. For brevity, table 5 displays the computing time in **Setting One** and do not report the corresponding results in another **Setting Two**.

Table 5. Computing time (seconds) using KNN-SVM classifiers and their competitors in **Setting One**.

		Kernels		
Classifiers	K	linear	polynomial	radial
CC-SVM	-	0.315	0.302	0.364
KNN-SVM	5	1.007	1.023	1.105
KNN-SVM	10	1.079	1.102	1.176

CC-SVM, KNN-SVM only using completely observed samples.

5. Discussions and Conclusion

In the paper, we develop new KNN-SVM classifiers in complex real data on disease diagnosis. Essentially, KNN-SVM classifiers are expansions of the well-known SVM with more "powerful functions" in dealing with complex imbalanced and missing data simultaneously. The reason why we call them KNN-SVM classifiers due to the following two aspects to accomplish their "powerful functions" using k-nearest neighbours' information: (1) Impute missing data by calculating k-nearest neighbours' statistical characteristic values and (2) interpolate some new samples between k-nearest minority class examples.

To further investigate KNN-SVM classifiers' performances in accuracies, the paper chooses three types of kernels and two K values in calculating the imputation values of missing data. Through the

real data, we find that classifiers with kernel "radial" always have relatively better performances in accuracy than the classifiers with other two kernels "linear" and "polynomial".

In the future, the paper will continue to develop more valuable diagnosis tools and methods in both static and dynamic settings. From the perspective of dealing with imbalanced data, more classifiers in our paper belong to the category of over-sampling methods. More specifically, the proposed KNN-SVM classifiers generate new minority class samples to obtain relatively equal numbers of samples in both majority and minority classes. In the future, we will carry out more investigations on developing classifiers with under-sampling techniques, which also reflect the idea of "k-nearest neighbours" and enrich our investigations in classifiers for more complex diagnosis.

Declaration of Competing Interest

The author declares that he has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Acknowledgments

The author is very grateful to all of the reviewers for their insightful comments and to the interviewees for participating in our investigation. The author's work was supported by the National Natural Science Foundation of China (72001197), National Statistical Science Research Project of National Bureau of Statistics (2021LY052), Innovation Centre for Digital Business and Capital Development of Beijing Technology and Business University (SZSK202317) and the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China (16XNH102). The author wants to thank his parents, his wife Yujie Liu and his cute babies: QQ and QD.

References

[1] Miroslav K, Holte R C Matwin S 1998 Machine Learning for the Detection of Oil Spills in Satellite Radar Images *Machine Learning* **30(2)** 195-215.
[2] Ivan T 1976 A Two Modifications of Cnn *IEEE Transactions on Systems, Man and Cybernetics* **6(11)** 769-772.
[3] Peter E H 1968 The Condensed Nearest Neighbor Rule *the IEEE Transactions on Information Theory* **14(3)** 515-516.
[4] Wilson L D 1972 Asymptotic Properties of Nearest Neighbor Rules Using Edited Data *IEEE Transactions on Systems, Man and Cybernetics* **2(3)** 408-421.
[5] Nitesh V C, Bowyer K W, Hall L O, et al. 2002 Smote: Synthetic Minority over-Sampling Technique *Journal of Artificial Intelligence Research* **16(3)** 321-357.
[6] Cheng H 2020 Comparison of partial least square algorithms in hierarchical latent variable model with missing data *Simulation: Transactions of the Society for Modeling and Simulation International* **96(10)** 825-839.
[7] Cortes C and Vapnik V. 1995 Support vector networks *Machine Learning* **20** 273-297.
[8] James G, Witten D, Hastie T, and Tibshirani R 2021 *An Introduction to Statistical Learning with Applications in R* Springer Texts in Statistics (STS)
[9] Little R J A and Rubin D B 1987 *Statistical Analysis with Missing Data* (New York: Wiley).
[10] Cheng H and Pei R M 2022 Visualization analysis of functional dynamic effects of globalization talent flow on international cooperation *Journal of Statistics and Information* **37(11)** 107-116.
[11] Wei C H, Wang S J and Su Y N 2022 Local GMM estimation in spatial varying coefficient geographocally weighted autoregressive model *Journal of Statistics and Information* **37(11)** 3-13.