# Development of Parkinson's Disease Classifier using SMOTE and Support Vector Machine

Haewon Byeon*
*Department of Medical Big Data, College of AI Convergence, Inje University
e-mail:byeon@inje.ac.kr

# SMOTE와 서포트 벡터 머신을 이용한 파킨슨병 분류기 개발

변해원*
*인제대학교 AI융합대학 메디컬 빅데이터전공

Abstract

This study minimized the imbalance issue by employing Synthetic Minority Over-sampling Technique (SMOTE), developed eight Support Vector Machine (SVM) models for predicting Parkinson’s disease using different kernel types ((C-SVM or Nu-SVM)×(Gaussian kernel, linear, polynomial, or sigmoid algorithm)), and compared the accuracy, sensitivity, and specificity of the developed models. This study evaluated 76 senior citizens with Parkinson's disease (32 males and 44 females) and 285 healthy senior citizens without Parkinson's disease (148 males and 137 females). The analysis results showed that the liner kernel-based Nu-SVM had the highest sensitivity (62.0%), specificity (81.6%), and overall accuracy (71.3%). The results of this study implied that developing a prediction model by using linear kernel-based Nu-SVM would be more accurate than other kernel-based SVM models for handling imbalanced disease data.

## 1. Introduction

Motor-symptoms (e.g., resting tremor, rigidity (slowing body movements down) are commonly observed in the early stage of Parkinson's disease [3]. Overthe past 20 years, many studies [5] have focused on nonmotor-symptoms such as autonomic nervous system dysfunction, dysesthesia, and cognitive impairment, which are observed in the early stages of Parkinson's disease. Shulman et al.(2001)[8] reported that these nonmotor-symptoms were found in 88% of Parkinson's disease patients. Patients with Parkinson's disease do not need any help in performing their daily activities in the early stages [9] because their symptoms can be well controlled with a small amount of medication. However, as Parkinson's disease progresses, since their cognitive and motor functions decline a lot, it becomes difficult to conduct their daily activities and eventually lose the ability to perform them independently [10]. As a result, they must rely on others [10]. In addition, diminished cognitive functions have been reported as a factor causing both the patient and the family to fall into despair and depression along with the gradual decline in Parkinson’s disease patients’ physical function and uncertainty about the progression of the disease [12]. Particularly, nonmotor-symptomsof Parkinson's disease such as cognitive impairment are major predictors for the morbidity of Parkinson's disease dementia. Therefore, it is necessary to detect them as soon as possible, whichrequires to accurately distinguish the cognitive decline in normal aging from thatin Parkinson's disease.

Many recent studies [21] have widely used support vector machine (SVM), a supervised learning algorithm, as a way to classify and predict complex risk factors of diseases. When developing a prediction model using binary data like a disease, it is highly likely to encounter an imbalanced issue because the number of patients is smaller than that of people without the disease [24]. The imbalanced issue may cause a prediction error in the process of conducting machine learning and degrade the performance of the model. Consequently, it needs an additional imbalanced data processing technique using sampling in order to resolve the prediction error due to the imbalanced data. Previous studies [26] have reported that synthetic minority over-sampling technique (SMOTE) has less overfitting than oversampling or undersampling. This study minimized the imbalance issue by employing SMOTE, developed eight SVM models for predicting Parkinson’s disease using different kernel types ((C-SVM or Nu-SVM)×(Gaussian kernel, linear, polynomial, or sigmoid algorithm)), and compared the accuracy, sensitivity, and specificity of the developed models.

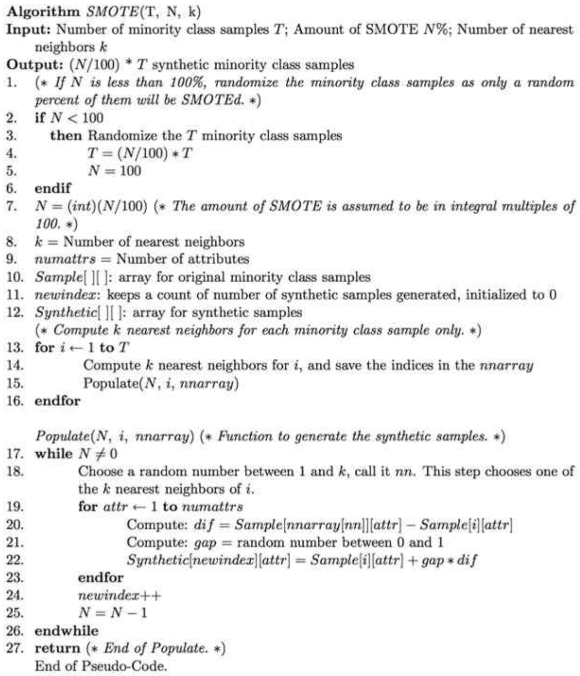
## 2. Methods

This study evaluated 76 senior citizens with Parkinson's disease (32 males and 44 females) and 285 healthy senior citizens without Parkinson's disease (148 males and 137 females) living in Seoul, Incheon, and Gwangju, while a senior citizen was defined as people equal to or older than 60 years and equal to or younger than 74 years. The power of this study was examined using G-Power version 3.1.9.7 (Universität Mannheim, Mannheim, Germany). The results showed that, when the number of predictors was 19, alpha=0.05, power (1-B)= 0.95, and the effect size (f2) was 0.15, the required number of samples was 217. Therefore, it was concluded that the number of this study’s samples (n=361) was enough to test statistical significance.

This study measured the cognitive levels for each subtype using the Cognition Scale for Older Adults (CSOA)[28], which could measure cognitive function comprehensively considering age and education level. CSOA is composed of eight subtests: Mini-Mental Status Examination in the Korean Version(MMSE-K), Verbal Memory Test, Stroop Test, General Information, Digit Span Test, Rey Complex Figure Test (RCFT), Confrontation Naming Test, and Verbal Fluency Test. This study transformed the raw scores of the eight subtests into standardized scores with a mean of 100 and a standard deviation of 15, and used them to develop prediction models. This study used SMOTE to over the imbalance issue of this binary dataset. SMOTE finds n nearest neighbors, belong to the same minor class, for any value of a minor class, draws a straight line with that neighbor, and creates random values until they show a synthetic ratio. SMOTE's algorithm is presented in Fig. 1.

## 3. Results

The analysis results showed that the liner kernel-based Nu-SVM had the highest sensitivity (62.0%), specificity (81.6%), and overall accuracy (71.3%). It was noteworthy that the polynomial-based C-SVM showed the highest specificity (86.5%) among the eight SVM models with the lowest sensitivity (28.8%). The linear kernel-based C-SVM had the lowest overall accuracy.



```
Algorithm SMOTE(T, N, k)
Input: Number of minority class samples $T$; Amount of SMOTE $N$%; Number of nearest
    neighbors $k$
Output: $(N/100) * T$ synthetic minority class samples
(* If N is less than 100%, randomize the minority class samples as only a random
percent of them will be SMOTEd. *)
if $N < 100$
  then Randomize the $T$ minority class samples
       $T = (N/100) * T$
       $N = 100$
endif
$N = (int)(N/100)$ (* The amount of SMOTE is assumed to be in integral multiples of
100. *)
$k$ = Number of nearest neighbors
$numattrs$ = Number of attributes
$Sample[\,][\,]$: array for original minority class samples
$newindex$: keeps a count of number of synthetic samples generated, initialized to 0
$Synthetic[\,][\,]$: array for synthetic samples
(* Compute k nearest neighbors for each minority class sample only. *)
for $i \leftarrow 1$ to $T$
        Compute $k$ nearest neighbors for $i$, and save the indices in the $nnarray$
        Populate($N$, $i$, $nnarray$)
endfor

Populate($N$, $i$, $nnarray$) (* Function to generate the synthetic samples. *)
while $N \neq 0$
        Choose a random number between 1 and $k$, call it $nn$. This step chooses one of
        the $k$ nearest neighbors of $i$.
        for $attr \leftarrow 1$ to $numattrs$
                Compute: $dif = Sample[nnarray[nn]][attr] - Sample[i][attr]$
                Compute: $gap$ = random number between 0 and 1
                $Synthetic[newindex][attr] = Sample[i][attr] + gap * dif$
        endfor
        $newindex$++
        $N = N - 1$
endwhile
return (* End of Populate. *)
End of Pseudo-Code.
```

[Fig. 1] Algorithm of SMOTE

## 4. Conclusions

In this study, the prediction accuracy of the linear kernel-based Nu-SVM algorithm was the highest when the prediction accuracy of the eight SVM classification algorithms was compared to evaluate the SVM performance by kernel type. The performance of nonlinear SVM is affected by the employed kernel function and the parameters constituting it [36]. The results of this study implied that developing a prediction model by using linear kernel-based Nu-SVM would be more accurate than other kernel-based SVM models for handling imbalanced disease data. Additional studies are needed to compare the accuracy using data from various fields to prove the prediction performance of linear kernel-based Nu-SVM..

## References

[1] C. R. Baumann, “Epidemiology, diagnosis and differential diagnosis in Parkinson's disease tremor”, Parkinsonism & Related Disorders, vol. 18, pp. S90-S92, 2012.

[2] K. Seppi, K. Ray Chaudhuri, M. Coelho, and S. H. Fox, R.

Katzenschlager, S. Perez Lloret, D. Weintraub, C. Sampaio, and the collaborators of the Parkinson's Disease Update on Non Motor Symptoms Study Group on behalf of the Movement Disorders Society Evidence Based Medicine Committee Search for more papers by this author, “Update on treatments for nonmotor symptoms of Parkinson's disease—an evidence based medicine review“, Movement Disorders, vol. 34, no. 2, pp. 180-198, 2019.

[3] L. M. Shulman, R. L. Taback, J. Bean, and W. J. Weiner, “Comorbidity of the nonmotor systoms of Parkinson's disease”, Movement Disordorders, vol. 16, no. 3, pp. 507-510, 2001.

[4] P. A. Koplas, H. B. Gans, M. P. Wisely, M. Kuchibhatla, T. M. Cutson, D. T. Gold, C. T. Taylor, and M. Schenkman, “Quality of life and Parkinson's disease”, Journal of Gerontology, vol. 54, no. 4, pp. M197–M202, 1999.

[5] D. A. Cahn, E. V. Sullivan, P. K. Shear, A. Prefferbaum, G. Heit, and G. Silverberg, “Differential contributions of cognitive and motor component processes to physical and instrumental activities of daily living in Parkinson's disease”, Archives of Clinical Neuropsychology, vol. 13, no. 7, pp. 575-583, 1998.

[6] A. J. Jones, R. G. Kuijer, L. Livingston, D. Myall, K. Horne, M. MacAskill, T. Pitcher, P. T. Barrett, T. J. Anderson, and J. C. Dalrymple-Alford, “Caregiver burden is increased in Parkinson’s disease with mild cognitive impairment (PD-MCI)“, Translational Neurodegeneration, vol. 6, no. 1, pp. 17, 2017.

[7] S. Lahmiri, and A. Shmuel, “Detection of Parkinson’s disease based on voice patterns ranking and optimized support vector machine”, Biomedical Signal Processing and Control, vol. 49, pp. 427-433, 2019.

[8] C. Jian, J. Gao, and Y. Ao, “A new sampling method for classifying imbalanced data based on support vector machine ensemble”, Neurocomputing, vol. 193, pp. 115-122, 2016.

[9] D. Elreedy, and A. F. Atiya, “A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance”, Information Sciences, vol. 505, pp. 32-64, 2019.

[10] H. K. Kim, and T. Y. Kim, Cognition Scale for Older Adults; CSOA: manual. Neuropsy Inc, Daegu, 2007.

[11] I. Steinwart, and A. Christmann, Support vector machines. Springer Science & Business Media, New York, 2008.