

Constructing a Sorting Machine for Degenerative Cerebrotopathia

Seohyun Moon*, Bogeum Choi*, Jiwoo An*, Taeseon Yoon**

*Natural Science, Hankuk Academy of Foreign Studies, Wangsan-ri, Mohyeon-myeon, Yongin-si, Gyeonggi-do Korea

**Faculty, Hankuk Academy of Foreign Studies, Wangsan-ri, Mohyeon-myeon, Yongin-si, Gyeonggi-do Korea

ski1999@hafs.hs.kr, bogeum8876@hafs.hs.kr, uiop5658@hafs.hs.kr, tsyoon@hafs.hs.kr

Abstract— Dementia with Lewy bodies and Parkinson's Disease has the common cause: mutation of gene CYP2D6. Thus, it is hard to distinguish those two diseases. In order to classify them clearly, construction of sorting machine is required. Various algorithms including Apriori, Decision tree, Support Vector Machine, K-means and Neural Network was used in the construction of sorting machine. Two kinds of sorting machines were constructed: one classifying whether the patient is under Dementia with lewy bodies or Parkinson's Disease, the other classifying whether the gene is normal or mutation has undergone.

Keywords— *Parkinson's Disease, Dementia with Lewy bodies, Classification Machine, Function Construction, Artificial Intelligence Algorithms*

I. INTRODUCTION

As human life span has expanded, the number of people suffering degenerative cerebrotopathia is continuously increasing. Most commonly occurred cerebrotopathia are dementia, and Parkinson's disease. Since degenerative cerebrotopathia may lead to failure of memory, and incapacities in normal life, detection of the disease is very important. Also, as symptoms of degenerative neural cerebrotopathia differ rather greatly or trivially, accurate detection of the disease and the cause is very important. Application of various algorithms in analyzation of complex disease such as neural disease has the profit of achieving fast, accurate detection in various angles without any errors. Thus, sorting machine of degenerative cerebrotopathia was constructing by using various algorithms for those reasons mentioned above.

II. RELATIVE RESEARCH

A. Parkinson's Disease

One of the common neurodegenerative disorders, Parkinson's disease (PD), is known to occur because of mutation in the alpha-synuclein gene. Property about this disease is, in substantia nigra, we can find degeneration of dopaminergic neurons.

B. Dementia with Lewy bodies

It is known as secondary common cause of neurodegenerative dementia in elders. It is kind of primary dysautonomia, which includes Parkinson's disease. All of these primary autonomic failures have same neurotic pathology, which is based upon synaptic protein α -synuclein's abnormal aggregation. The specific symptoms of DLB differs from other kinds of dementia, but similar to Parkinson's disease.

C. Apriori Algorithm

Apriori is an algorithm that figures out associations between data by checking frequencies. It is one of the major types of algorithm based on association rule. The major purpose of usage of Apriori algorithm in data mining is to find out pattern in data set. Apriori algorithm calculates conditional probability of each case and returns the most probable set of data, which appears most frequently in input data.

In Apriori algorithm, 'support' means the number of transaction in item set. Apriori algorithm runs through data and figures out support value of each itemset. Any itemset that has support value above standard minimum support value is called as large itemset. Rest of itemsets that has support value below standard minimum support value is called small itemsets. Only large itemsets are passed to the next classification stage to create candidate itemsets. Same process applies to candidate itemsets and it is repeated until only one large itemset is left.

D. Decision Tree Algorithm

A decision tree is an algorithm used to classify data. Classification is conducted by mapping attribute-value vectors (instances) in a class to classes based on their attribute values. This form of mapping is called the decision tree. As the aim of the decision tree is to predict the most probable class of the instance, its searching process can be divided into two methods depending on the type of the tree. Trees can be divided into two types, a type only with a leaf node, and the other with a test node. A tree with a test node undergoes classification process several times until a subtree consisting

the tree is connected to the result, which is leaf node. It is proved mathematically that smaller trees show higher predictive accuracy. The process of classifying could also be seen as a process of minimizing entropy of instances when instances are regarded as objects. At the root of the tree, the entropy of the instances is maximum, but through the process of dividing and conquering, entropy reduction is maximized and as a result, instances are classified successfully.

E. Support Vector Machine

Support vector machine is a programming for recognizing pattern and analysis of data. It can be categorized as linear and nonlinear. First, Linear is that classifying data which is suggested beforehand with line in hyperplane. Hyperplane is third-dimensional space that is realize input-output pairs (x, y) in which $y \in Y$ is interchangeable and variable. For example, it can be used for graph, lattices. In similarity, non-linear apply polynomial function to classify data. SVM is most appropriate when most nearby hyperplane between classified data is farthest from classes. A line for classifying data is just required existing hyperplane between data, however farthest line from each data is most appropriate model. We could use the very SVM to find similarity between amino-acid, consisting amino mutation inversely, and using KH test to analysis how feasible mutation that is.

F. K-means

K-means signifies that classify data using cluster. If the number of data is 'N', it will be classified by cluster designating 'K' which is less than the number of 'N'. Each of data in cluster has similarity with another data in cluster which is differ from others by dissimilarity. It minimizes inefficient cost function. The way adjusting center of cluster contains three process. Firstly, Selecting K data object of central point in data object set D. and secondly, the central point section of K data object have other data object divided due to distance which central point is more nearby it. And finally, adjusting the center K data object until it convergences.

When cluster is gather about three for appearance, it is to turn inappropriate conclusion if user inputs below or more than number of K. Due to possibility of error in input number 'K', user need caution.

G. Neural Network

Artificial neural network is a programming which is resembled with the central nervous of human body. It solves problems by adjusting control bond strength, and that resembled with the central nervous of human body for storing knowledge, and using for recognizing. It consists of input data, hidden layer, and output data. When input data enters though the hidden layer, artificial neural network directs process of signal segment for classification, is to derive in output layer data. Neural network has Supervised Learning which is based on Machine Learning which has vector type analogy what is

right outcome. It also materializes Unsupervised Learning is has difference with not having target value for input data.

III.RESULT

A. Apriori

Based on table 1, 2, 3, it can be analyzed that amino acid Leucine is most commonly occurred in the data of Alzheimer's and Parkinson's disease regardless of class, window, and position. Thus, in building a machine for classification of Alzheimer's and Parkinson's disease, using Leucine associated rules as a classification standard in high level could lead to lower accuracy result.

TABLE 1. BEST RULE IN 5 WINDOW

	amino1	amino2	amino4
x1	L		
x2	L		
x3	L		
x6			L
x7			L
x8	L		
x9		L	
x10			L

TABLE 2. BEST RULE IN 7 WINDOW

	amino1	amino3	amino4
x1		L	
x2		L	
x3			L
x5		L	
x6	L		
x7	L		
x8			L

TABLE 3. BEST RULE IN 9 WINDOW

	amino4	amino6	amino7
x1			L
x2			L
x3	L		
x5			L
x6		L	
x7	G		
x8	L		
x9			L
x10	G		

B. Decision Tree

From the data attained by using decision tree algorithm, some repeated combinations of amino acids in each folder could be found. By extracting such patterns, power of influence in each position could be compared. In other words, it is possible to determine which position is in higher classification standard by analyzing data attained through decision tree algorithm. Table 4,5,6 below is shows

comparison of the level of classification standard between positions in each class combination occurrences. In this experiment, data was analyzed in three different sizes: 5 window, 7 window, 9 window. '>' in table 4,5,6 indicates that a position in the left of '>' is in higher classification standard

TABLE 4. RULE EXTRACTION FROM 5 WINDOW

Class	Position	Amino Acid
1,8	pos1>pos2	A
3,8	pos2>pos5	F
7,8	pos2>pos3	G
7,8	pos2>pos1	Y
1,1	pos1>pos5	W
2,8	pos5>pos4	D
3,5,7,8	pos5>pos1	P
3,7,8	pos5>pos2	E
5,1	pos5>pos1	F
7,8	pos5>pos2	S
2,4	pos4>pos2	T
7,8	pos4>pos1	F
3,8	pos4>pos1	H
3,5,7,8	pos4>pos3	P
3,8	pos4>pos1	Q
3,8	pos4>pos3	L
3,7,8	pos4>pos1	A
7,8	pos4>pos5	R
7,8	pos4>pos1	F
7,8	pos4>pos5	E
3,8	pos4>pos1	Q
3,6,7,8	pos4>pos2	L
7,8	pos4>pos1	H
3,8	pos4>pos1	Q
1,2,5,8	pos4>pos2	T
7,8	pos4>pos5	D
3,8	pos2>pos3	G
3,7,8	pos2>pos4	P
3,8	pos2>pos5	F
7,8	pos2>pos1	Y
2,3	pos4>pos3	V
3,7	pos4>pos1	Q
3,8	pos4>pos2	I
5,7	pos4>pos3	P
7,8	pos4>pos1	H

TABLE 5. RULE EXTRACTION FROM 7 WINDOW

Class	Position	Amino Acid
1,3,6	pos5>pos1	G
1,3,	pos3>pos1	R
3,7	pos3>pos1	R
3,8	pos7>pos3	L
3,8	pos3>pos5	H
3,6,7	pos3>pos1	R
4,8	pos3>pos7	V
4,8	pos3>pos6	T
4,5,7,8	pos3>pos7	V
5,7,8	pos2>pos7	E
5,7,8	pos5>pos6	F
7,8	pos3>pos6	G
6,8	pos2>pos3	W

TABLE 6. RULE EXTRACTION FROM 7 WINDOW

Class	Position	Amino Acid
3,5	pos7>pos5	M
3,5	pos7>pos9	F
3,6	pos4>pos8	L
5,9	pos7>pos6	M
5,7	pos4>pos5	G
4,5	pos4>pos2	V
5,6	pos4>pos5	G
5,8	pos4>pos7	R
5,8	pos6>pos3	G
5,6,7	pos4>pos3	G

Based on the data from table 4, 5 and 6, influence of each position in data classification process could be compared between class combinations.

TABLE 7. COMPARISON OF INFLUENCE OF EACH POSITION IN CLASS COMBINATIONS

Window	class	position comparison
5 window	1,8	pos4>pos5>pos2>pos1
	2,3	pos4>pos3
	2,4	pos4>pos2
	2,8	pos5>pos4>pos2
	3,7	pos4>pos3
	3,8	pos4>pos2>pos3>pos5>pos1
	5,7	pos4>pos3
	7,8	pos4>pos5>pos2>po1
7 window	3,7	pos3>pos1 pos3>pos5 (pos1 and pos5 cannot be compared based on the data)
	3,8	pos7>pos3>pos5
	4,8	pos3>pos6 pos3>pos7 (pos6 and pos7 cannot be compared based on the data)
	5,7	pos3>pos7 pos2>pos7 pos5>pos6 (pos2 and pos3 cannot be compared based on the data)
	6,8	pos2>pos3
	3,6	pos5>pos1 pos3>pos1 (pos5 and pos3 cannot be compared based on the data)
9 window	3,5	pos7>pos5 pos7>pos9 (pos5 and pos9 cannot be compared based on the data)
	3,6	pos4>pos8

C. SVM

From SVM, we knew that we need non-linear function to classify the class. Since we have to classify 9 classes, it's natural for us to get non-linear function.

D. K-means

Although we tested 9 classes, each class is just one kind of Parkinson's or Alzheimer's disease. Therefore, we clustered into two groups only to classify Parkinson's and Alzheimer's disease. In the table 8, SSE (Cluster Sum of Squared error) means "how far" the data is from the regression line. The smaller the SSE is, the better this function can classify the data. Moreover, from here, we can find out window 9 is best at classifying data as we have thought. In addition, big cleavage and SSE shows that even though the data is in same cluster, the arrange of cluster is large that it has may include wrong data or classify wrong when we put the new data set.

TABLE 8. K-MEANS SSE, CLEAVAGE

	Cluster Sum of Squared error	Cleavage (full data)	Cleavage (Cluster 0)	Cleavage (Cluster 1)
window 5	3505.0	1	7	8
window 7	3316.0	1	1	2
window 9	3237.0	1	1	2

E. Neural Network

As you can see in table 9, we fail to classify the correct class in Neural Network. But when we divide this classes into two groups, Parkinson's and Alzheimer's Disease, we can find that accuracy of data has noticeably grown up. Also, we found that class 1,2,4 / class 3,7 / class 4,6,8,9 has high rate of similarity. From this, we find out that some Alzheimer's Disease from different cause or some Parkinson's Disease from different cause is hard to classify compared to just classifying Parkinson's and Alzheimer's Disease. We thought the reason of this is we got definite rules between them. And by choosing the suitable window value, the accuracy can get more higher.

TABLE 9. ACCURACY OF NEURAL NETWORK CLASSIFICATION IN CLASS LEVEL AND GROUP LEVEL

	Class (%)	Group (%)
window 5	3.2303	59.2697
window 7	1.5717	97.6424
window 9	2.2613	96.9849

F. Process of experiment

We use program that are decision tree, apriori, svm, K-means and neural network for classifying Alzheimer disease dementia of lewy bodies disease. To find out what is the most contain amino acid, we use apriori algorithm. It represents that the most frequency amino acid, second most frequency amino acid, and third most frequency acid. Decision tree classifies position which contains similar pattern of amino acid for each fold. After that, the most frequency combination acid is selected main position of the fold which is in it. And then user deduces each position of window distinguish with another window. Svm is to represent a point of similarity with standard. However, this case dissimilarity is so high, we deduces non-linear is more appropriate within it. To classify

data figure between entry section and cluster, we use K-means. For distinguishing Alzheimer and dementia of lewy bodies, input value of K should be two. Neural network is used for distinguishing class data which is revealed by percent of accuracy. It used for classifying between class in Alzheimer, class in dementia of lewy bodies, and between Alzheimer and dementia of lewy bodies.

IV. CONCLUSIONS

To synthesize all of the algorithms we used, we concluded that to make a good sorting machine which can classify even reason, we need non-linear function and satisfy the rules we extracted from the decision tree at least. In this case, it can get high error rate due to large similarity between some Alzheimer's disease and Parkinson's disease from different cause. On the other hand, classifying without each reason can bring good result just using neural network. To get higher accuracy and use less time, which is good classifying machine, we can use k-means first to find out where the SSE gets bigger and use Neural Network considering the SSE of k-means.

We expect this machine can help doctors predict or diagnose the exact disease patient got.

REFERENCES

- [1] S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.
- [2] J. Breckling, Ed., *The Analysis of Directional Time Series: Applications to Wind Speed and Direction*, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, 1989, vol. 61.
- [3] S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.
- [4] M. Wegmuller, J. P. von der Weid, P. Oberson, and N. Gisin, "High resolution fiber distributed measurements with coherent OFDR," in *Proc. ECOC'00*, 2000, paper 11.3.4, p. 109.
- [5] R. E. Sorace, V. S. Reinhardt, and S. A. Vaughn, "High-speed digital-to-RF converter," U.S. Patent 5 668 842, Sept. 16, 1997.
- [6] (2002) The IEEE website. [Online]. Available: <http://www.ieee.org/>
- [7] *FLEXChip Signal Processor (MC68175/D)*, Motorola, 1996.
- [8] "PDCA12-70 data sheet," Opto Speed SA, Mezzovico, Switzerland.
- [9] A. Karnik, "Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP," M. Eng. thesis, Indian Institute of Science, Bangalore, India, Jan. 1999.
- [10] J. Padhye, V. Firoiu, and D. Towsley, "A stochastic model of TCP Reno congestion avoidance and control," Univ. of Massachusetts, Amherst, MA, CMPSCI Tech. Rep. 99-02, 1999.
- [11] *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification*, IEEE Std. 802.11, 1997.
- [12] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.
- [13] R. J. Vidmar, On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. 21 (3). pp. 876–880. Available: <http://www.halcyon.com/pub/journals/21ps03-vidmar>, Aug. 1992.
- [14] H. Braak and E. Braak, "Neuropathological Staging of Alzheimer-related Changes." SpringerLink. Springer-Verlag, Acta Neuropathol 17 May 1991.



Seohyun Moon was born in Busan, Korea, in 1999. Since 2015, she has been studying in the Department of Natural Science, Hankuk Academy of Foreign Studies. Her research interest is in artificial intelligence, neuroscience, and virtual reality.



Bogeum Choi was born in Daegu, Korea, in 1999. Since 2015, she has been studying in the Department of Natural Science, Hankuk Academy of Foreign Studies. Her research interest is in artificial intelligence and neuroscience.



Jiwoo An was born in Yeong ju, Korea, in 1999. Since 2015, she has been studying in the Department of Natural Science, Hankuk Academy of Foreign Studies. Her research interest is in artificial intelligence and bioinformatics.



Taeseon Yoon was born in Seoul, Korea, in 1972. He received the Ph.D. candidate degree in computer education from the Korea University, Seoul, Korea, in 2003. From 1998 to 2003, he was with EJB analyst and SCJP. From 2003 to 2004, he joined the Department of Computer Education, University of Korea, as a lecturer and Ansan University, as an adjunct professor. Since December 2004, he has been with the Hankuk Academy of Foreign Studies, where he was a computer science and statistics teacher. He was the recipient of the Best Teacher Award of the Science Conference, Gyeonggi-Do, Korea, 2013.