Feature Selection for The Classification of Clinical Data of Stroke Patients

Onny Setyawati[1], Aji Seto Arifianto[2], and Moechammad Sarosa[3]
[1]Brawijaya University, Jl. MT Haryono 167 Malang 65145, Indonesia
[2]Politeknik Negeri Jember, Jember, Indonesia
[3]Politeknik Negeri Malang, Malang, Indonesia
Email: osetyawati@ub.ac.id

***Abstract*--Clinical examination of the patient with suspected stroke to determine the type of pathology is still widely applied, especially in Indonesia due to constraints in the implementation of the Gold Standard Procedure. Clinically, the examination of the various features starts from the physical symptoms, medical history and laboratory results, which might take long duration and costly. Moreover, not all inspection features have a significant influence to distinguish the type of stroke, hence, sorting features are required. The selection process to get the best features is performed by identifying similarity to the features of each class. Fuzzy Entropy generates the entropy value from the degree of membership of each feature. The result of the implementation of feature selection is able to select 13 of the best features with 96% in accuracy, therefore, the process is more effective than having to check 32 features.**

***Index Terms*--infarction, hemorrhagic, feature selection, fuzzy entropy.**

I. Introduction

Determination of stroke by pathological classification on the new stroke patients is urgent, due to the need for the prevention of blockage (infarction) or bleeding (hemorrhagic) continuously. An analytic descriptive study was conducted to determine the clinical symptoms dominant in determining the diagnosis model of stroke. Based on the fact that for bleeding stroke predominant clinical symptoms include decreased consciousness, headache, systolic blood pressure of more than 179 and vomiting. While in ischemic stroke no clinical symptoms were dominant [1]. Applying the Gold Standard for diagnosis of stroke in Indonesia in particular areas encountered several obstacles, including the patient's condition unallowed to move, high cost, longer duration and the risk of radiation effects [2]. Alternative early diagnosis of stroke can be obtained by the clinical examination, including asking the perceived symptoms of the patient (anamnesis), data retrieval, patients medical history and a neurological examination [3], the process produce patient data variable called features. According to Verikas some features are redundant or even irrelevant [4] so it is necessary to do the screening. Features used in this study were 32, consisting of patients clinical data of physical examination, medical history and laboratory test results. This screening allows the diagnosis of stroke disease process to run faster and more accurate, especially supported by using computer technology as a tool.

Duen-Yian Yeh et al conducted a study in Taiwan with the classification of optimization techniques of cerebrovascular disease prediction, by using attributes as much as 29 consisting of disease diagnosis data, physical examination data, and numerical data from blood tests. Decision tree method was used to get the best results with an accuracy of 99.59% [5].

Pasi Luukka argued if the classification feature selection in the case was important, since it could reduce computational costs, reduce noise and improve the accuracy of the classification process. This study used fuzzy entropy with the results for the Parkinson data of 85.03% by only using two features of 22 features, and reached 98.28% for the dermatology with 29 features of 34 features [6]. Research on the determination of the dominant features of tuberculosis has been done by Beni Widiawan et al. Their study described the diagnosis of tuberculosis disease through the examination of sputum (phlegm), and it showed that not all containing germs tuberculosis sputum. The process of feature selection methods used Fuzzy Entropy (FE) and Fuzzy Entropy-Normalization (FE-N), and the Max-Min-Redundancy Relevance (mrmr), whereas for the classification they used Backpropagation Neural Networks. The result reached 94.98% for FE, 95.80% for FE-N, and 86.03% for mrmr, and FE showed faster process than the mrmr [7].

This study proposed a new idea for the diagnosis of stroke based on the clinical data of patients using Fuzzy Entropy for the feature selection and Learning Vector Quantization Neural Network (LVQ-NN) method to determine the success of the process [6][9].

II. Methodology

The data of stroke patients from two hospitals were used [6]. After the data were collected, the normalization took place to unify the data of each feature. Data that was not a number need to be converted into numerical data. For instance, diabetes history data "1" for yes and "0" for no. Data for gender generally in the form of textual data both men and women, which will be converted into "male = 1" and "female = 0". Furthermore, all data features will be normalized in the same range of between [0,1] [8].

Table 1 shows some examples of the normalization of data, where a diverse range of numbers turned into a value between [0] and [1]. The data in column diagnosis were only converted into numerical data since it described the outcomes, i.e. for stroke Infarction = "1" and

978-1-5386-3246-8/17/$31.00 © 2017 IEEE

Hemorrhagic stroke = "2". Next, the data were divided into two for the training and testing data.

TABLE I SAMPLE RESULTS OF NORMALIZATION DATA

AGE		SYSTOLIC	
REAL	NORMALIZATION	REAL	NORMALIZATION
42	0,172413793	120	0,25
58	0,448275862	180	0,5
55	0,396551724	156	0,4
70	0,655172414	180	0,5

Feature based training data were selected to look for more dominant features compared to the others using the Fuzzy Entropy (FE). The FE methods took a sample of data from each column of the dataset and then divided it into several classes using the similarity classifier [7]. Similarity classifier is calculated by Yu's norms.

If the calculation produced unideal class, then it obtained similarity value of zero, however, when including the ideal class, then the value obtained similarity of 1. By using the fuzzy entropy value μA(xj), entropy value would be higher if the similarity value approachs 0.5, means that it is increasingly heterogeneous and vice versa.

III. RESULTS AND DISCUSSION

The manual calculation of Fuzzy Entropy is shown in Table II. The data consisted of three features, i.e. Hb, Chlo and HDL, each having 4 pieces of data. From these data the Ideal vector can be derived from the average data of each feature of each class. Ideal vector for class 1 Hb features is shown in the following (Table III).

The results of the calculation of all the similarity $S(x, v)$ is shown in Table IV.

TABLE II DATA FROM MANUAL CALCULATION OF FE

DATA	HB	CHLO	HDL	CLASS
P1	0.516	1	0.0739	1
P2	0.559	1	0.0595	1
P3	0.451	0.523	1	2
P4	0.527	0.569	0.972	2

TABLE III IDEAL VECTOR

IDEAL VEC	HB	CHLO	HDL
1	0.5375	1	0.0667
2	0.489	0.546	0.986

TABLE IV RESULTS OF SIMILARITY VALUE S(X, V)

$S(x, v)$	HB	CHLO	HDL
P1,A	1	1	1
P2,A	1	1	1
P3,A	1	0,523	0,067
P4,A	1	0,569	0,999
P1,B	1	0,546	0,089
P2,B	1	0,546	0,074
P3,B	1	1	0,986
P4,B	1	1	1

After that the value of Fuzzy Entropy can be determined. Calculating the value of fuzzy entropy was preceded by calculating the similarity of each feature with the existing classes. Output of feature selection is classified for patients clinical data features. The levels are based on the value of FE, top rated are considered as the most relevant features.

Table V shows the level (ranking) features of FE calculation results with 323 learning data.

Validation results of the feature selection is achieved by comparing between the results of research based on the opinion of experts (medical doctors) and the previous research which has been the benchmark of medical world, especially in neurology for the classification of stroke.

Three experts confirmed that the blood pressure (systolic, distol) and awareness were the most important features distinguishing infarction and hemorrhagic stroke. Testing was performed to see the impact on the performance of the classification of the stroke feature selection. 50 experiments used different data from learning data to confirm the objectivity and diagnosis result of the neurologist. It aims to determine the accuracy of program outcomes by referring to the unbias diagnosis. The tests used a learning rate (α) of 0.5.

Table VI presents the results of some experiments with a number of different features. It shows that with 32 features 70% in accuracy was obtained for Hemorrhagic class (35 data were correct and 15 were incorrect). Significant improvement occurred with 19 to 13 features where the accuracy of 96% was achieved.

TABLE V LEVEL (RANKED) FEATURES BASED ON FE

No	Name of Feature	FE Value	No	Name of Feature	FE Value
1	Hemoglobin	19,6298	17	Age	104,122
2	Chlorida	20,8683	18	Calcium	106,887
3	Hdl (High-Density Lipoproteins)	22,4922	19	Calium	116,434
4	Natrium	23,0310	20	Seizures	121,733
5	Breathing	33,3624	21	Temperature	240,713
6	Heartbeat	46,3342	22	Cholesterol	242,151
7	Diastole	49,0162	23	Cardiovascular	300,254
8	Total_Cholesterol	54,9201	24	Diabetes	366,745
9	Uric_Acid	59,0912	25	Motion Limited	368,293
10	Creatinine	59,2385	26	Difficulty Talk	382,343
11	Systole	59,5995	27	Nauseous_Vomit	388,448
12	Ldl (*Low*-Density Lipoproteins)	65,0567	28	Hypertension	403,095
13	Awareness	67,5414	29	Albumin	411,034
14	Triglyeride	76,8386	30	Headache	415,421
15	Kga (Glucose)	95,6672	31	Gender	442,924
16	Blood Urea Nitrogen	101,764	32	Body_Weakness	445,213

TABLE VI EXPERIMENTAL RESULTS AND TIMING

NUMBER OF FEATURE	CORRECT DATA	TIMING (SEC)
32	70%	0,004489
20	72%	0,008319
19	96%	0,005156
13	96%	0,002182
10	70%	0,002157

Detailed mapping of the validation test results can be seen in Table VII.

TABLE VII

		REAL	
		INFARC	HEMORRHAGIC
PROGRAM	INFARC	35	0
	HEMORRHAGIC	0	15
TOTAL		35	15

Furthermore, to verify the results of the diagnostic tests included in the category of excellent, very good or good or even fail measurements, ROC curve and the result of Area Under The ROC curve = 0.500 (*red arrow*) helped to explained it. ROC curve with 32 features is shown in Figure 1. ROC with 13 features is in Figure 2, the result of Area Under The ROC curve for 13 features = 0.967, this means that the test conditions was classified in **excellent (A)**.

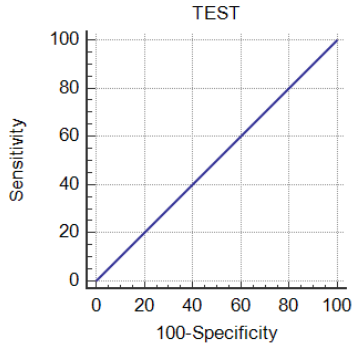

Fig. 1. ROC curve with 32 features.

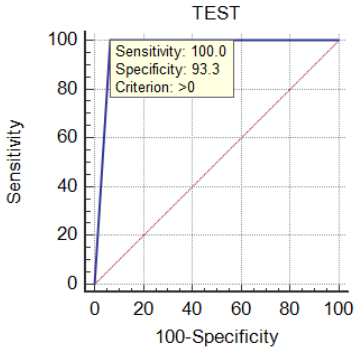

Fig. 2. ROC curve with 13 features.

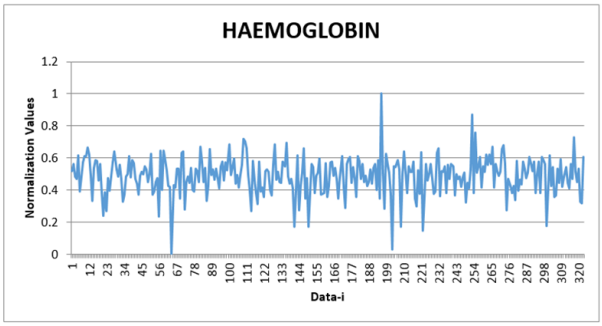

Fig. 3. Graph data distribution Haemoglobin features.

TABLE VIII

	AWARENESS	
STROKE	STABLE (0)	DECREASE (1)
INFARC	227	4
HEMORRHAGIC	3	89

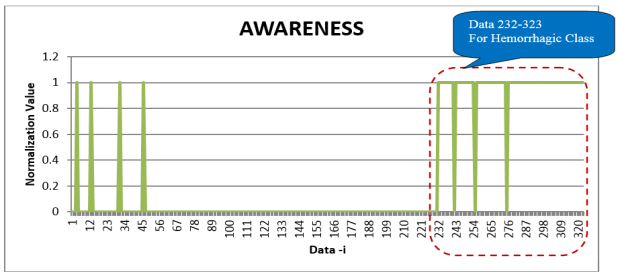

Fig. 4. Graph data distribution Awareness features

Three distribution data led to differences in the levels after the calculation of Fuzzy Entropy. Examples of three different distribution of the data is represented with the features of haemoglobin, awareness and body weakness in the following.

In Figure 3 the learning dataset haemoglobin is in the range of 0.4 to 0.6 for all classes. Haemoglobin had a similarity value close to 1 that led to its smallest FE value and was considered to have the best relevance, however, this was not enough to distinguish the classes.

The second example for the distribution of data is the awareness, as shown in Table VIII. There is a fundamental difference to the value of the class of hemorrhagic and infarction. In majority there´s no decrease in the awareness of infarction class, meanwhile the Hemorrhagic class was dominated by the condition of decreased awareness. These values were considered as the most ideal for distinguishing the classes, although the awareness was on rank 13. Figure 4 shows a graph of the distribution of the data for awareness feature.

As for the features of weakness there were only two possibilities (0) and (1) of the value of learning impartial dataset (Figure 5), indicating no classes were identical with the condition of the body weakness. The similarity value approached 0.5 that caused high FE value, and this reflected the level of relevance that this feature was a bad class. Similar to the haemoglobin, this value could not be used as a reference to distinguish classes.

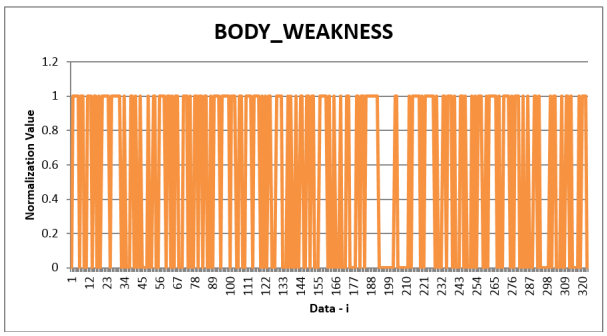

Fig. 5. Graph data distribution Body-weakness features

IV. Conclusions

Feature selection using Fuzzy Entropy method managed to reduce the number of features for the classification of 32 to 13 features, and to accelerate the computing time (computation time of 0.166213 seconds was obtained). However, Fuzzy Entropy has the disadvantage that it does not pay attention to the relationship between the features of the others. If further work will be conducted on the similar case, it can use the methods of classification or other feature selection which especially consider the relationship between features and classification methods.

References

[1] M. Bahrudin, "Pengertian Stroke.Pdf," *J. Saintika Med.*, vol. 6, no. 13, 2010.

[2] A. C. Widjaja, "Uji diagnostik pemeriksaan kadar D-dimer plasma pada diagnosis stroke iskemik," Semarang, 2010.

[3] M. Bahrudin, "Bahrudin_M_2009.pdf," *J. Saintika Med.*, vol. 5, no. 11, 2009.

[4] A. Verikas and M. Bacauskiene, "Feature selection with neural networks," *Pattern Recognit. Lett.*, vol. 23, no. 11, pp. 1323–1335, 2002.

[5] D.-Y. Yeh, C.-H. Cheng, and Y.-W. Chen, "A predictive model for cerebrovascular disease using data mining," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8970–8977, 2011.

[6] P. Luukka, "Feature selection using fuzzy entropy measures with similarity classifier," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4600–4607, 2011.

[7] B. Widiawan, I. K. E. Purnama, M. H. Purnomo, J. T. Elektro, I. Teknologi, and S. Nopember, "Penentuan Fitur Dominan Pada Penyakit Tuberkulosis Berbasis Fuzzy Entropy," in *14th Seminar on Intelligent Technology and its Application (SITIA2013)*, 2013, pp. 299–302.

[8] O. P. Master and S. T. Stockholm, "Implementing LVQ for Age Classification Implementing LVQ for Age Classification," 2007.

[9] A. S. Arifianto, M. Sarosa, and O. Setyawati, "Klasifikasi Stroke Berdasarkan Kelainan Patologis dengan Learning Vector Quantiation," *Eeccis*, vol. 8, no. 2, pp. 117–122, 2014.

[10] C. Iyakaremye, P. Luukka, and D. Koloseni, "Feature selection using Yu's similarity measure and fuzzy entropy measures," in *IEEE International Conference on Fuzzy Systems*, 2012.

[11] M. Cebrián, M. Alfonseca, and A. Ortega, "The normalized compression distance is resistant to noise," *IEEE Trans. Inf. Theory*, vol. 53, no. 5, pp. 1895–1900, 2007.