INVITED PAPER
3rd Placer, 2013 ISI Jan Tinbergen Award for
Outstanding Young Statistician

Classification of Congenital Hypothyroidism using Artificial Neural Networks

Iris Ivy M. Gauran
University of the Philippines Diliman

Ma. Sofia Criselda A. Poblador
University of the Philippines Manila

The Newborn Screening Reference Center (NSRC) of the National Institutes of Health in the University of the Philippines Manila collects measurements from five attributes to determine whether Congenital Hypothyroidism (CH) is present in a neonate. Detecting the CH cases is a major concern of medical practitioners because it provides richer information than the healthy ones. However, because of the rarity of this metabolic condition, existing classification algorithms oftentimes misclassify a newborn as "normal" even if it is not. This paper investigates the efficiency of Self-Organizing Kohonen Maps (SOM), a type of artificial neural network. Though it is a visualization and clustering tool, the researchers want to probe on its ability to detect outliers and properly classify a newborn as normal or not by coming up with a statistically computed threshold value. Instead of working directly with the original attributes of the data, a reduced set of SOM prototypes is utilized to represent the data in a space of smaller dimension, seeking to preserve the probability distribution and topology of the input space. Results showed a misclassification rate of 13.5%. Though it is found to be slightly less superior to the existing classification rules, the proposed methodology was able to address the problem of finding a statistical threshold value. Also, the methodology verifies that age has a major effect on misclassifying "Normal" as "Abnormal" since postponement of newborn screening to a later age causes the quantization error to boost drastically, hence, easily exceeding the value of the first decision threshold.

***Keywords**: Self-Organizing Kohonen Maps (SOM), classification algorithm, outlier detection, newborn screening for congenital hypothyroidism*

1. Introduction

In the advent of the internet era, browsing for the number of Filipinos born per hour would lead any researcher to a lot of websites, most of which declaring different and inconsistent values. Indeed, tracking such phenomena is not an easy task because of the richness of the data available everywhere. In particular, the volume of newborn screening datasets of certain hospitals poses significant problems for biochemists, medical doctors and other professionals handling the said data.

In the Philippines, the newborn screening program started in 1996 when a group of pediatricians and obstetricians from 24 hospitals in the Metropolitan Manila area collaborated to establish the incidence of six metabolic conditions – congenital hypothyroidism, congenital adrenal hyperplasia, galactosemia, phenylketonuria, homocystinuria and glucose-6-phosphate dehydrogenase deficiency (Padilla, 2003). Since most babies with metabolic disorders look normal at birth, a couple will never know that their baby has the disorder until the onset of signs and symptoms. At this point, it is often observed that the ill effects (mental retardation and even death) are already irreversible.

Consequently, the Department of Health (DOH) has recognized the significance of the initial data and efforts are now being undertaken to ensure the nationwide implementation of neonatal screening. In lieu with this, the need for data mining algorithms to uncover the subtle patterns that may be hidden in these medical datasets also arises. Data mining offers a suite of algorithms, each addressing a different task and in the process elucidating a unique facet of the data (Fayad et al., 1996). Of the many facets of data mining, the researchers are mainly interested in using Artificial Neural Networks (ANNs) in solving a classification problem.

Exploratory data mining using ANNs offers an alternative dimension to data mining since they have a natural propensity to learn; they learn how to solve problems from data as opposed to solving problems based on explicit problem specification (Craven and Shavlik, 1997). Also, ANNs provide tremendous opportunities pertaining to data classification because the learning characteristics of ANNs enable them to deal efficiently with noisy data – partial, possibly incorrect and potentially conflicting data (Lu et al., 1996).

In this paper, the researchers aim to probe into the efficiency of Self-Organizing Kohonen Maps (SOM), a type of artificial neural network, as an unsupervised classification technique. The proposed methodology will be applied to a one-week newborn screening data collected by the Newborn Screening Reference Center (NSRC) of the National Institutes of Health (NIH) in University of the Philippines Manila. In one-week's time, a total of 5,843 newborns underwent a screening test for congenital hypothyroidism, which entailed measurements from five distinct attributes. An analysis of these attributes using the proposed methodology will eventually lead to the classification of a newborn as either "Normal" or

"Abnormal." In both practice and literature, the amount of thyroid-stimulating hormone (TSH) found in the blood sample serves as the deciding factor for the classification of newborns. An elevated TSH value indicates an impaired functioning of the thyroid gland, thus a newborn with a high value of TSH is deemed "abnormal" and is highly likely to develop congenital hypothyroidism.

Congenital hypothyroidism (CH) results from an abnormality in thyroid gland development and is one of the most common preventable causes of mental retardation. Worldwide, it is known that the incidence of congenital hypothyroidism is 1:4,000. Thus, in the given data set, it is expected that there are no more than ten cases of CH observed. This scenario is highly susceptible to misclassification errors since the proportion of CH cases is relatively low compared to that of healthy newborns. In addition, health professionals are more concerned with the CH cases because it provides much more valuable information than the healthy cases. Though SOM is a widely popular visualization and clustering tool, the researchers will be delving on the possibility of using an unsupervised learning technique, such as SOM in classifying data patterns. Amidst the existence of well-performing classification models, this paper aims to evaluate the performance of SOM particularly in the classification of rare classes. The facts that most of the medical datasets contain rare cases introduce additional challenges to the classification algorithm. Hence, this paper aims to provide a comparison between the implementation of rule-based classification using SOM and the rule-based classification algorithm readily available in Weka. More importantly, this paper aims to provide a novel technique in addressing the lack of statistically-computed threshold value for determining cases as either "normal" or "abnormal."

2. The Self-Organizing Map

The self-organizing map (SOM) is a type of artificial neural network which is an excellent tool in exploratory phase of data mining. It projects input data on prototypes of a low-dimensional regular grid that can be effectively utilized to visualize and explore properties of the data. These prototypes are sometimes called nodes or neurons. A typical SOM network consists of two layers of nodes, an input and output layer. Each node in the input layer is fully connected to the nodes in the output layer. The number of input nodes corresponds to the number of input variables while the number of output nodes usually depends on the given problem and is specified by the user. The key concept in training SOM is the neighborhood around a winning node, which is the collection of all nodes with the same radial distance.

At the start of the training process, weights are initially assigned to each node in the output layer and a neighborhood size large enough to cover half of the neurons in the output layer is also chosen. When an input pattern, randomly chosen from the training set is presented to the network, each node in the output layer calculates how similar the input is to its initial weights. The similarity is

often measured by some distance between the input pattern and the weight of the neuron. The neuron with the minimum distance is the winning neuron and its weight, as well as the weights of its neighboring neurons are strengthened or updated to be closer to the value of the input pattern. The goal of SOM training is to represent all data points from a high-dimensional space into a two-dimensional feature map, without altering the distance and proximity of the points from each other. Therefore, the training of SOM is unsupervised and competitive with the winner-takes-all strategy.

The algorithm maps the input data to a two-dimensional feature map containing the output neurons with the following steps and repeats the overall loop several times until there are no significant changes in the weight vectors of the output neurons.

(1) An input vector $\underline{x}$ randomly chosen from the training set is compared with all the output nodes i with an initial weight vector $\underline{w}_i$ and the best matching unit (BMU) on the map is identified. The BMU is the node k whose weight vector $\underline{w}_k$ has the smallest Euclidean distance from the input vector. That is,

$$||\underline{x} - \underline{w}_k|| = \min_i ||\underline{x} - \underline{w}_i||$$

(2) The weight vector of the best matching unit k as well as those of the nodes in the neighborhood of k are updated so as to move towards the current input pattern $\underline{x}$ using the Kohonen learning rule:

$$\underline{w}_i^{new} = \underline{w}_i^{old} + h_{ik}\,(\underline{x} - \underline{w}_i^{new}), \text{ if } i \text{ is in the neighborhood of the BMU } k$$

$$\underline{w}_i^{new} = \underline{w}_i^{old}, \text{ if } i \text{ is not in the neighborhood of the BMU } k$$

where h_{ik} is the neighborhood kernel function, which is a decreasing function of the distance between the i^{th} and k^{th} weight vectors on the feature map. A widely used neighborhood function is based on the Gaussian function is given by

$$h_{ik} = \alpha \exp\left\{-\frac{\|r_i - r_k\|^2}{2\sigma^2}\right\}$$

where $0 < \alpha < 1$ is the learning rate factor, r_i and r_k are the positions of neurons i and k on the SOM grid and σ is the neighborhood radius, which is also decreasing monotonically.

(3) Repeat steps 1 and 2 until all input patterns from the training set are processed. To achieve a better convergence towards the desired mapping, it

is usually required to repeat the previous loop until weights are stabilized, while decreasing the size of the neighborhood.

Once the SOM training algorithm has converged, the computed feature map displays important statistical characteristics of the input space, which can be summarized as follows (Gonçalves et al., 2011):

i. *Vector Quantization*: the basic objective of SOM is to store a large set of input vectors by finding a smaller set of prototypes that provides a good approximation of the input space.

ii. *Topological Ordering*: the feature map computed by SOM is ordered topologically. Similar input vectors are mapped close to each other, while dissimilar ones are mapped far apart.

iii. *Density matching*: the SOM reflects the probability distribution of data in the input space. Regions in the input space in which the input patterns are taken with a high probability of occurrence are mapped onto larger domains of the output space, and thus have better resolution than regions in the output space from which input patterns are taken with a low probability of occurrence.

3. The Proposed Methodology

The methodology proposed in this work attempts to exploit the characteristics and properties of the SOM to perform classification of a newborn according to a screening test result for a certain harmful or potentially fatal disorder. The key point of the proposed method is to perform the classification of a newborn through a set of SOM prototypes instead of working directly with the original attributes of the data. The SOM is used to map the original patterns of the data to a reduced set of prototypes arranged in a two-dimensional rectangular grid. The objective is to represent the data in a space of smaller dimension, seeking to preserve the probability distribution and topology of the input space.

In this processing level, two important factors should be considered: the sampling process of the input data and the determination of the SOM training parameters. The following parameters and criteria to train SOM will be utilized by the proposed methodology:

(1) For the sampling process, the input data will be randomly divided into ten subsamples. Of these ten randomly partitioned subsets, one subsample will be used as a validation set and the remaining nine subsamples will be used as the training set.

(2) Deboeck and Kohonen recommend using ten times the dimension of the input patterns as the number of neurons, and this will be adopted in the methodology. The two-dimensional feature map will be constructed with a map size of *15 x 15*, with a total of 225 output neurons.

(3) Input neurons will have randomly initialized weights. The weights are derived from the Uniform Distribution multiplied by the corresponding range of values of each of the attributes.

(4) The accuracy of the map also depends on the number of iterations of the SOM algorithm. A rule of thumb states, for good statistical accuracy, the number of iterations should be at least 500 times the number of neurons. In order to achieve sufficient training, the total number of iterations will be 1,000 and 1,500 for the whole learning process.

(5) Learning rate is initially set to 0.75. There is no guideline in suggesting good learning rates to any given learning problem. In standard SOM, too large and too small learning rates can lead to poor network performance. However, some researchers posit that the learning rate is associated with the number of iterations. To maintain a balance between those two researcher-specified elements of the algorithm, given that the value of the learning rate is 0.75, 1,000 and 1,500 iterations were both considered.

(6) Neighborhood radius is initially set to 7.5. A larger neighborhood radius is usually used in the beginning of the training which gradually decreases to a suitable final radius. Past literature suggests that the initial radius size be set to half the size of the feature map.

Once the Self-Organizing Map has been trained, quantization errors ε_q ($\underline{x}$, $\underline{w}_k$) will be calculated to obtain a measure of proximity between the input vector $\underline{x}$ and the weight vector of its best matching node $\underline{w}_k$ from the feature map. A set of quantization errors $\{\varepsilon_q^{\,i}\}$ will be generated for each training vector $\underline{x}_i$ where i = 1,2,..., N, using the formula for Euclidean distance, $\varepsilon_q^{\,i}(\underline{x}_i, \underline{w}_k^{\,i}) = \|\underline{x}_i - \underline{w}_k^{\,i}\|$ where $\underline{w}_k^{\,i}$ is the best matching unit of training vector $\underline{x}_i$.

Given the set of computed quantization errors, a classification rule could be developed by forming a decision threshold τ^+. Past studies constructed a decision threshold based on the 95th percentile of the distribution of the quantization errors of the training vectors. In a nutshell, this particular single decision threshold resolves the problem of classification by simply applying the principle of goodness-of-fit test. That is, for an unseen record $\underline{x}_{new}$, the following statements could be regarded as the null and alternative hypotheses, respectively: *Ho*: $\underline{x}_{new}$ is classified as "Normal" vs. *Ha*: $\underline{x}_{new}$ is classified as "Abnormal"

The decision rule is to reject *Ho* if $\varepsilon_q^{\,new}(\underline{x}_{new}, \underline{w}_k^{\,new}) \geq \tau^+$. Thus, if the computed quantization error of the new training vector exceeds the decision threshold τ^+, then $\underline{x}_{new}$ will be classified as "Abnormal."

The training data used during the learning process of SOM only consists of the "normal" state of the system being monitored. In this light, any unseen record

whose quantization error is considerably large would be most likely classified as "abnormal."

An important modification of this paper seeks to address the problem of classifying rare data patterns by introducing the use of binary, instead of single decision threshold. In particular for positively-skewed data, the rare classes would presumably fall on the extreme-positive portion of the distribution. Yet, there could be few "abnormal" data patterns which would lie not too far from the bulk of the input space and hence might be misclassified as "normal."

The methodology proposes that it is not enough to compare the quantization error ε_q^{new} of a new input pattern $\underline{x}_{new}$ with the 95th percentile (τ^+) of the distribution of the quantization errors of the training vectors. A comparison with a measure of dispersion, specifically, the interquartile range (IQR_{ε_q}) of the distribution of the quantization errors of the training vectors would verify if the new input pattern $\underline{x}_{new}$ is significantly "normal" or "abnormal."

Based on the classification rule, if ε_q^{new} exceeds τ^+, the 95th percentile of the distribution of quantization errors of the training vectors, then $\underline{x}_{new}$ will be classified as "abnormal." However, if the contrary is true, then the proposed binary decision threshold is suggesting to compare ε_q^{new} with IQR_{ε_q} first before concluding a decision. If ε_q^{new} exceeds IQR_{ε_q}, then $\underline{x}_{new}$ will be classified as "normal but needs repeat measurement." Otherwise, if ε_q^{new} does not exceed IQR_{ε_q}, then $\underline{x}_{new}$ will be classified as "normal."

4. Results and Discussion

There are three existing problems which the health professionals involved in newborn screening would be most interested in solving: (i) identify a suitable threshold value for TS1 (ii) investigate the significance of utilizing age in the classification and (iii) classify a baby as "normal" or "abnormal." Due to the lack of statistical researches in this particular field, health professionals usually rely on their expertise and knowledge in order to determine the answers to the abovementioned problems.

Given the data on newborn screening for Congenital Hypothyroidism with five distinct attributes, a *15 x 15* lattice structure was constructed for SOM training. The parameters and criteria discussed in the previous chapter were applied during the learning process. After the construction of feature map, quantization errors were computed for each training vector. From the distribution of the quantization errors, the 95th percentile, which is the first decision threshold τ^+, is found to be 48.0176. Subsequent comparison of the quantization error of each validation sample against the first decision threshold was done. Only the 26th observation from the validation set exceeded 48.0176. Thus, the 26th observation from the validation set is classified as "abnormal." However, its actual classification is "normal" with TS1, TS2, and TS3 values of 0.81. It is of paramount importance that the said baby is already 87 days old when it took the test. Since the baby took

the newborn screening way outside the prescribed 24-48 hour period, the outlying value of the variable age affected the choice of the best matching unit for the data point.

Medical experts theorize that babies who took the screening test after a month or so, would tend to be categorized as "abnormal" because the time elapsed could already be spent treating the child had s/he been classified as "abnormal." Generally, it can be observed that babies who were screened more than a month after they were born tend to be misclassified by the algorithm. This has some serious implications for parents who took the screening test for granted. Because of a myriad of factors such as money, marital concerns, inaccessibility of the test and so on, these parents did not make sure that their child should take the screening test on the appropriate time. Thus, they are subjected to some amount of scare because their child was classified as "abnormal" even though s/he is actually "normal."

To assess the accuracy of the proposed binary decision thresholds, each of the validation sample was then fed into the trained feature map with the corresponding quantization errors computed. The quantization errors of the validation sample were compared against the first decision threshold and subsequently, against the interquartile range, both obtained from the training sample. Each observation from the validation sample was classified according to the proposed binary classification rule.

Table 1. Misclassified "*Normal*" Babies in the Validation Sample Using the Proposed Binary Classification Rule

OBS	CLASS	ASSIGNMENT	AGE	TS1	TS2	TS3	TS99
22	Normal	Retest	34	2.17	2.17	2.17	2.17
23	Normal	Abnormal	47	1.4	1.4	1.4	1.4
24	Normal	Abnormal	58	1.51	1.51	1.51	1.51
25	Normal	Abnormal	60	1	1	1	1
26	Normal	Abnormal	87	0.81	0.81	0.81	0.81

Observations 31 to 56 are misclassified as "normal" instead of "retest" because of the instrument/technical error in the value of the TS1. This means that the quantization errors of these observations fell in below the value of the interquartile range which is 7.120. Likewise, these cases are considered anomalous because of the large deviation between the TS1, TS2 and TS3 values. In practice, the values of these three variables are not supposed to deviate that much from each other so these observations that have doubtful TS1 values led to misclassification.

Table 2. Misclassified "*Retest*" Babies in the Validation Sample Using the Proposed Binary Classification Rule

OBS	CLASS	ASSIGNMENT	AGE	TS1	TS2	TS3	TS99
31	RETEST	NORMAL	1	11.18	4.49	5.22	4.86
33	RETEST	NORMAL	1	11.28	8.12	8.52	8.32
37	RETEST	NORMAL	1	10.83	9.86	9.59	9.72
39	RETEST	NORMAL	1	12.22	8.75	9.33	9.04
40	RETEST	NORMAL	1	11.44	8.74	9.44	9.09
41	RETEST	NORMAL	2	10.5	8.46	9.73	9.09
43	RETEST	NORMAL	2	10.3	9.19	9.01	9.1
44	RETEST	NORMAL	2	10.66	10.57	8.46	9.52
46	RETEST	NORMAL	2	11.03	8.81	10.62	9.72
48	RETEST	NORMAL	2	10.02	10.2	9.95	10.08
49	RETEST	NORMAL	3	10.38	6.74	5.9	6.32
50	RETEST	NORMAL	3	10.37	9.87	10.09	9.98
51	RETEST	NORMAL	3	10.65	7.32	7.61	7.46
52	RETEST	NORMAL	3	10.95	10.36	8.99	9.67
53	RETEST	NORMAL	3	10.11	7.26	7.11	7.18
54	RETEST	NORMAL	4	10.76	8.6	8.56	8.58
56	RETEST	NORMAL	5	11.7	7.11	7.78	7.44

Finally, observations 64, 67, 70, 71 and 72 are misclassified as "retest" even though they are supposed to be "abnormal." One explanation for this is the TS1, TS2 and TS3 values which are more or less in the borderline between the values of "retest" and "abnormal."

Table 3. Misclassified "*Abnormal*" Babies in the Validation Sample Using the Proposed Binary Classification

OBS	CLASS	ASSIGNMENT	AGE	TS1	TS2	TS3	TS99
64	ABNORMAL	RETEST	1	14.14	13.55	10.92	15
67	ABNORMAL	RETEST	1	12.96	11.13	13.02	15
70	ABNORMAL	RETEST	2	14.96	13.71	13.79	15
71	ABNORMAL	RETEST	2	12.74	14.08	12.8	15
72	ABNORMAL	RETEST	9	14.04	14.2	12.56	15

At this point, it can be observed that the three major concerns of the medical experts are adequately addressed by the proposed methodology. On the other hand, the interpreted Weka outputs fail to answer two of the three problems previously mentioned. In fact the only concern that can be directly answered by the algorithms in Weka is the classification of the data points.

Using the ZeroR classifier, it can be observed that all values that are supposed to be "retest" and "abnormal" are misclassified as "normal." In other words, all observations are classified in just one class ("normal") because of the very small number of outlying observations found in the dataset. This limitation of classification is highly talked of in different literatures.

Another classifier used is the JRip algorithm. It performs better than ZeroR because there are only three misclassified observations compared to the 46 misclassified observations in ZeroR.

Hence, if we compare the percentage of correctly classified instances, we can say that the outputs of Weka are better than that of the proposed methodology. However, the researchers still claim that the proposed methodology is still superior in addressing real-life data problems which is not just classification, that is, when seeking the patterns presented by the empirical evidence.

Furthermore, besides looking at the number of misclassified observations, we should also look into the impacts of the misclassification. That is, the outputs of the algorithms in Weka would tend to misclassify a baby into a "normal" state given that necessary treatments should already be given immediately. Conversely, the misclassifications in the proposed methodology are not considered as grave mistakes in actuality because a child considered as "retest" rather than "abnormal" does not pose serious problems unlike that of "abnormal" which is classified as "normal."

5. Summary and Conclusion

In this paper, the Kohonen's self-organizing feature map (SOM) is utilized in grouping newborns with more or less similar characteristics. The SOM is then used for classification by computing quantization errors and comparing each with a binary decision threshold. A classification rule using a binary decision threshold is proposed:

1st decision threshold τ^+: If $\varepsilon_q^{new}(\underline{x}_{new}, \underline{w}_k^{new}) \geq \tau^+$

then $\underline{x}_{new}$ is classified as "Abnormal."

2nd decision threshold $\text{IQR}_{\varepsilon_q}$: If $\varepsilon_q^{new}(\underline{x}_{new}, \underline{w}_k^{new}) < \tau^+$ then

(1) If $\varepsilon_q^{new}(\underline{x}_{new}, \underline{w}_k^{new}) \geq \text{IQR}_{\varepsilon_q}$ then $\underline{x}_{new}$ is classified as "Normal but needs Repeat Measurement" or "Retest."

(2) If $\varepsilon_q^{new}(\underline{x}_{new}, \underline{w}_k^{new}) < \text{IQR}_{\varepsilon_q}$ then $\underline{x}_{new}$ is classified as "Normal."

The proposed classification rule is then compared with classification algorithms from WEKA. A misclassification rate of 13.5% is found to be slightly less superior to the existing classification rules. However, the proposed methodology was able to address the problem of finding a statistical threshold for TS1 as a reasonable and sound cut-off value. In addition, the methodology employed by this paper was able to verify that age has a major effect on misclassifying “Normal” as “Abnormal." It was found out that older babies are more likely to be misclassified. This is due to the fact that the variable age is causing the quantization error to boost drastically, hence, easily exceeding the value of the first decision threshold.

REFERENCES

CRAVEN, M. and SHAVLIK, J.,1997, Using neural networks for data mining, *Future Generation Computer Systems* 13(2-3):211-229.

FAYYAD, U., PIATETSKY-SHAPIRO, G., SMYTH, P., UTHURUSAMY, R., 1996, *Advances in Knowledge Discovery and Data Mining*, Menlo Park:CA:AAAI Press.

GONCALVES, M.L., COSTA, J.A.F. and NETTO, M.L.A., 2011, Land-cover classification using self-organizing maps clustered with spectral and spatial information, In Mwasiagi, J, (ed.) *Self Organizing Maps-Applications and Novel Algorithm Design*, Croatia: InTech.

LU, H., SEITIONO, R., and LIU, H., 1996, Effective data mining using neural networks, *IEEE Transactions on Knowledge and Data Engineering* 8(6):957-961.

PADILLA, C.D., 2003, Newborn screening in the Philippines, *Southeast Asian J Trop Med Public Health* 34 Suppl 3:87-8.