

Comparison of AI Systems in Fetal Health Classification

Krzysztof Brzoza^a, Kamila Gendasz^a and Michał Dzumla^a

^a Faculty of Applied Mathematics, Silesian University of Technology, Kaszubska 23, 44100 Gliwice, POLAND

Abstract

Fetal health is a very important aspect in this day and age. The ability to predict the health status based on preliminary examinations could help a lot in saving lives, especially if it is done by an artificial intelligence system, which would streamline the whole process. The issue is complicated by the multitude of different algorithms that vary in accuracy and execution time that can be used for such predictions. This project compares three popular artificial intelligence systems: k-nearest neighbor algorithm, naive Bayes classifier, soft sets to show which one of them will prove to be the best. The final results are shown in comparison graphs. In this article, the mathematical issues related to these algorithms and examples of their application in this task are presented.

Keywords

soft sets, knn, naive Bayes

1. Introduction

The reduction of child mortality is one of the key indicators of human progress. It is expected that by 2030, countries end preventable deaths of newborns and children under 5 years of age, with all countries aiming to reduce under-5 mortality. Parallel to the notion of child mortality is of course maternal mortality, during and following pregnancy and childbirth. The vast majority of these deaths occurred in low-resource settings, and most could have been prevented.

Cardiotocographs (CTGs) is a simple and cost-accessible option to assess fetal health, allowing healthcare professionals to take action in order to prevent child and maternal mortality. The equipment itself works by sending ultrasound pulses and reading its response, thus shedding light on fetal heart rate (FHR), fetal movements, uterine contractions, and more.

Artificial intelligence is very important in all in all the aspects of modern life [1, 2, 3]. In medical areas it is used for the prognosis of some diseases, classification, and even clustering. It can be seen in the research in this area. In [4, 5], the neural network were used for prediction covid-19 virus spread. Similar research was the analysis of mortality [6]. Again in [7, 8], the deep learning solutions were analyzed and discussed in terms of application in medical image analysis tasks. Image analysis was also used in the skin evaluation mechanism [9]. In medical systems not only neural networks are used but also fuzzy logic and computational intelligence. In [10], the fuzzy sets approach was used in the evaluation of the health-

care problems. All mentioned solutions use different approaches but the incoming data are analyzed mainly in two ways - taking the data, or only the extracted features. Such solutions were analyzed and discussed in terms of data analysis in [11, 12, 13, 14, 15].

This work will be compared artificial intelligence systems, their performance, and accuracy in predicting fetal health.

For this purpose, 3 artificial intelligence systems were used:

- The k-nearest neighbours clustering algorithm [16].
- Naive Bayes classifier [17].
- Softsets [18].

to check which one of them will obtain the highest accuracy.

2. The mathematical part of the selected tool

- The k-nearest-neighbor clustering algorithm, to maximize the probability of finding the most suitable solution, searches for it among k most similar solutions, and then chooses the most popular one by voting.

The easiest way to determine the similarity is to use the distance, the smaller it is, the greater the similarity is. In this case, was used Minkowski metric [19].

$$M_m(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^m \right)^{1/m} \quad (1)$$

where

- x_i is the value from the sample vector.

SYSTEM 2021 @ Scholar's Yearly Symposium of Technology, Engineering and Mathematics. July 27–29, 2021, Catania, IT

✉ krzybrz165@student.polsl.pl (K. Brzoza);

kamigen389@student.polsl.pl (K. Gendasz);

michdzu822@student.polsl.pl (M. Dzumla)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

Table 1

Sample value

id	baseline value	...	hist._tend
1	120.0	...	1.0

Table 2

Used data

id	base. val.	...	his._tend	fet._health
1	120.0	...	1.0	2.0
2	132.0	...	0.0	1.0
...
2126	140.0	...	1.0	2.0
2127	142.0	...	0.0	1.0

- y_j is the value of the currently selected vector from the training set.
- i index of the currently analysed column from the database.

Description of the KNN clustering algorithm:

1. Performing data normalization.
2. Counting the distance between the test vector and all vectors of the training set.
3. Sorting distances.
4. Selection of the label of the k-nearest vectors in relation to the test vector.
5. Assign the most frequent label as the label of the test vector.
6. In the event of a tie in the voting, the selection shall be made by a lot.

To improve the performance of the algorithm, data normalization was applied. This causes all dimensions for which the distance is calculated to have equal significance. Otherwise, a situation could arise in which a single dimension would dominate other dimensions. Thanks to normalization we obtain a situation, where the values of a variable belonging to the interval [0;1].

$$x_j(i) = \frac{x_j(i) - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (2)$$

- i is another index of the vector.
- j is the index of the variable.
- $\max(x_j)$ is the maximum value of j .
- $\min(x_j)$ is the minimum value of j .

2.1. A calculation example for our base

Assuming that our sample vector is shown in Tab. 2.

And having a training set, the Minkowski metric was used to calculate the distance between the sample vector and the analyzed vectors of the training set:

$$M_2(x, y) = \left(\sum_{i=1}^{21} |x_i - y_i|^2 \right)^{1/2} \quad (3)$$

where x_i is the value from the column of the sample vector, and y_i is the value of the column of the currently analyzed vector from the training set. In project was assumed the value $m = 2$. After analyzing the entire training set, the distances between the sample vector and the training set vectors are sorted. Next, a neighborhood is selected for the specified k nearest neighbors based on the previously calculated distances, and the most frequent label is selected.

A naive Bayes classifier based on Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4)$$

is particularly suitable for problems with very many dimensions in the input. Despite the simplicity of the method, it often performs better than other very complex classifier methods. The classifier learns by analyzing a set of learning data for which the correct classes are given.

The model includes $P(X|Y)$ -the probability of observations for different class labels and $P(Y)$ -a prior probability, which is the probability calculated before the random experiment is performed. This is the classical probability calculated in the same way as the overall probability. Bayes' rule is used to determine $P(Y|X)$ -the conditional probability of a class for a given observation. The label for which the probability is highest is chosen. The model is called naive because it assumes a very strong simplification that says that for a fixed class label all are features are independent of each other.

2.1.1. A calculation example for this base

Classify the object $X = (133.0, \dots, 0.0)$ for this table (decision means 1.0 - Normal, 2.0 - Suspect, 3.0 - Pathological) what is shown in Tab. 2.

After assigning a column to a 1.0/2.0/3.0 group, it is needed to determine which probability is higher.

- $P(C = C_1|X)$ e.g. $P(\text{decision} = 1.0 | \text{base. val.} = 133.0, \dots, \text{hist._tend} = 0.0)$,
- $P(C = C_2|X)$ e.g. $P(\text{decision} = 2.0 | \text{base. val.} = 133.0, \dots, \text{hist._tend} = 0.0)$,
- $P(C = C_3|X)$ e.g. $P(\text{decision} = 3.0 | \text{base. val.} = 133.0, \dots, \text{hist._tend} = 0.0)$

So there is the need to calculate $P(C = C_1) \cdot P(C = C_1|X)$, $P(C = C_2) \cdot P(C = C_2|X)$ and $P(C = C_3) \cdot P(C = C_3|X)$ and compare the results.

1. $P(C = C_j) = \frac{1}{3}$
2. $P(a_i = [value]|C = C_j) = [probability]$ where under i insert the number of the next column, under $[value]$ the value of that column, and under $[probability]$ the calculated probability value, $j \in \{1, 2, 3\}$.

Then it is needed to substitute the calculated values into the formula:

$$P(C = C_j) \cdot \prod_{i=1}^{21} P(a_i = x_i|C = C_j) = [result_j] \quad (5)$$

It is checked which result (probability value) is greater which determines which class the object X should be classified in.

- Soft sets are defined by a set of properties which, however, are not completely precise (e.g. instead of saying that it is 15° C outside one can say that it is warm). The general theory of soft sets is very similar to that of fuzzy sets but much simplified. Thanks to the fact that there are no restrictions on the approximate description of the object, it is very easy to apply this method in practice. Mathematical assumptions: U - the class of all elements considered. E - parameter class $P(U)$ - set of all subsets of U $A \subseteq E$ - components considered. So the pair (F, A) is called a soft set, where F is an assignment:

$$F : A \rightarrow P(U) \quad (6)$$

Calculation example:

- U - fetal health class
- E - a set of parameters that determine the health of the fetus
- $A = \{\text{baseline value}, \dots, \text{hist.}_\text{tend.}\}$

The class consists of three health states.

$U = \{\text{Normal}, \text{Suspect}, \text{Pathological}\}$

$E = \{e_1, \dots, e_{21}\}$

What follows is the definition of subsets specifying the characteristics that health states satisfy given assumptions e_i , where $i \in 1 : 21$:

$F(e_1) = \{\text{Normal}\}$

$F(e_2) = \{\text{Normal}, \text{Suspect}\}$

...

$F(e_{20}) = \{\text{Pathological}\}$

$F(e_{21}) = \{\text{Normal}\}$

Table 3
Used data

U	baseline value	...	his._tend.
h_1	1	...	0
h_2	0	...	1
h_3	1	...	1

Based on this, a soft set table can be created (see Tab. 3).

Due to the simplicity of this system, subsets of a soft set can be determined very easily. E.g. (F, A) is a subset of (G, B) if $A \subset B$ and $\forall e \in A_1$ and $F(e)$ and $G(e)$ are the same. Two soft sets will be equal when the previous poset condition is satisfied and when (G, B) is also a poset of (F, A) .

3. Description of the database

The project was used a ready-made database, which can be downloaded at [20].

The database consists of one table with 22 columns. 21 columns are numerical data defining the CTG study and the last text column defines the fetal status.

Description of the columns:

- 'baseline value' - FHR baseline (beats per minute).
- 'accelerations' - Number of accelerations per second.
- 'fetal_movement' - Number of fetal movements per second.
- 'uterine_contractions' - Number of uterine contractions per second.
- 'light_decelerations' - Number of light decelerations per second.
- 'severe_decelerations' - Number of severe decelerations per second.
- 'prolonged_decelerations' - Number of prolonged decelerations per second.
- 'abnormal_short_term_variability' - Percentage of time with abnormal short term variability.
- 'mean_value_of_short_term_variability' - Mean value of short term variability.
- 'percentage_of_time_with_abnormal_long_term_variability' - Percentage of time with abnormal long term variability.
- 'mean_value_of_long_term_variability' - Mean value of long term variability.
- 'histogram_width' - Width of FHR histogram
- 'histogram_min' - Minimum (low frequency) of FHR histogram.

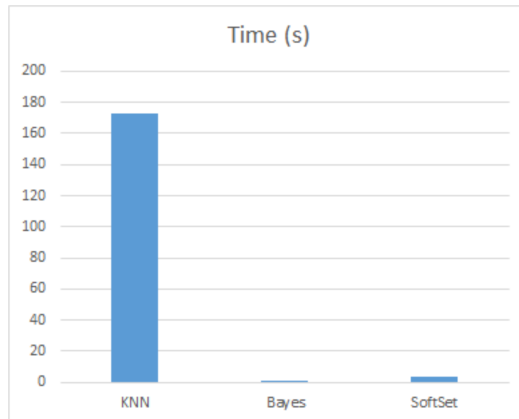


Figure 1: Time analysis.

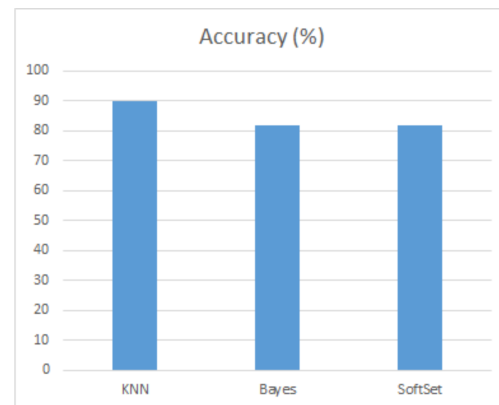


Figure 2: Obtained accuracy level.

- 'histogram_max' - Maximum (high frequency) of FHR histogram.
- 'histogram_number_of_peaks' - Number of histogram peaks.
- 'histogram_number_of_zeroes' - Number of histogram zeros.
- 'histogram_mode' - Histogram mode.
- 'histogram_mean' - Histogram mean.
- 'histogram_median' - Histogram median.
- 'histogram_variance' - Histogram variance.
- 'histogram_tendency' - Histogram tendency.

4. Tests

10 tests were performed for each system and extracted the average values, obtaining the following results:

- A k-nearest neighbours clustering algorithm:
 - for $k = 1$: 90.61%
 - for $k = 2$: 90.24%
 - for $k = 3$: 90.39%
 - for $k = 3$: 90.39%
- Naive Bayes-nearest-neighbour classifier: 82.43%
- Soft sets: 82.77%

5. Experiments

For each of them, a certain number of tests were performed and extracted average values to better specify the final result.

In the k-nearest neighbor clustering algorithm using the Minkowski metric, 10 tests were performed each for different k , where k was assumed to be 1, 2, 3, and 4.

For all k achieved average precision values were around 90, which is considered to be a very good result, and it is concluded that this algorithm performs very well for the database. Unfortunately, a big disadvantage for this algorithm is the execution time against a large database. In each iteration, for each validation sample, distances to nearest neighbors are recalculated and sorted anew, hence the execution time for the database, for one iteration was about 173 seconds.

In the naive Bayes classifier algorithm, the final results are not as satisfactory as in the k-nearest neighbor clustering algorithm. The average precision obtained for 10 iterations was about 82%. This is not a bad result, but compared to the previous algorithm, a significant difference can be seen. It was also checked if the algorithm would improve its final result when splitting the data 80:20 for training and validation data. Unfortunately, the precision did not improve. The advantage of this algorithm is the speed of its execution, oscillating around 0.3 seconds.

For classification using soft sets, an average obtained accuracy was oscillating 82%. The execution time of the algorithm for 1 iteration was about 4 seconds. In this method, quarterlies were used due to the strongly overlapping characteristics of the CTG surveys analyzed.

Graphs comparing the accuracy and execution time of the algorithms used in our project are shown in Fig. 1 and 2.

6. Conclusions

After an accuracy comparison, the k-nearest-neighbor clustering algorithm proved to be the best. However, its execution time was incomparably longer than in the other cases. So if we look at the overall optimality of the performance, the Bayes algorithm with a slightly worse

result for a much faster execution time can be determined as the best algorithm for our task.

However, the most important factor is the accuracy of the algorithm because the life of the fetus depends on the correctness of the diagnosis.

References

- [1] G. Capizzi, G. Lo Sciuto, C. Napoli, E. Tramontana, M. Woźniak, A novel neural networks-based texture image processing algorithm for orange defects classification, *International Journal of Computer Science and Applications* 13 (2016) 45–60.
- [2] G. Capizzi, G. Lo Sciuto, C. Napoli, R. Shikler, M. Woźniak, Optimizing the organic solar cell manufacturing process by means of afm measurements and neural networks, *Energies* 11 (2018). doi:10.3390/en11051221.
- [3] G. Lo Sciuto, G. Capizzi, S. Coco, R. Shikler, Geometric shape optimization of organic solar cells for efficiency enhancement by neural networks, *Lecture Notes in Mechanical Engineering* (2017) 789–796. doi:10.1007/978-3-319-45781-9_79.
- [4] M. Wiecek, J. Siłka, D. Połap, M. Woźniak, R. Damaševičius, Real-time neural network based predictor for cov19 virus spread, *Plos one* 15 (2020) e0243189.
- [5] S. I. Illari, S. Russo, R. Avanzato, C. Napoli, A cloud-oriented architecture for the remote assessment and follow-up of hospitalized patients, in: *Symposium for Young Scientists in Technology, Engineering and Mathematics*, volume 2694, 2020.
- [6] Z. Zhang, W. Yao, Y. Wang, C. Long, X. Fu, Wuhan and hubei covid-19 mortality analysis reveals the critical role of timely supply of medical resources, *Journal of Infection* 81 (2020) 147–178.
- [7] D. Karimi, H. Dou, S. K. Warfield, A. Gholipour, Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis, *Medical Image Analysis* 65 (2020) 101759.
- [8] S. Stolte, R. Fang, A survey on medical image analysis in diabetic retinopathy, *Medical image analysis* 64 (2020) 101742.
- [9] D. Połap, Analysis of skin marks through the use of intelligent things, *IEEE Access* 7 (2019) 149355–149363.
- [10] A. Mardani, R. E. Hooker, S. Ozkul, S. Yifan, M. Nilashi, H. Z. Sabzi, G. C. Fei, Application of decision making and fuzzy sets theory to evaluate the healthcare and medical problems: a review of three decades of research with recent developments, *Expert Systems with Applications* 137 (2019) 202–231.
- [11] C. Napoli, G. Pappalardo, E. Tramontana, An agent-driven semantical identifier using radial basis neural networks and reinforcement learning, in: *CEUR Workshop Proceedings*, volume 1260, 2014.
- [12] H. Das, B. Naik, H. Behera, Medical disease analysis using neuro-fuzzy with feature extraction model for classification, *Informatics in Medicine Unlocked* 18 (2020) 100288.
- [13] G. Capizzi, S. Coco, G. L. Sciuto, C. Napoli, A new iterative fir filter design approach using a gaussian approximation, *IEEE Signal Processing Letters* 25 (2018) 1615–1619.
- [14] D. Połap, M. Woźniak, Meta-heuristic as manager in federated learning approaches for image processing purposes, *Applied Soft Computing* 113 (2021) 107872.
- [15] S. Russo, S. I. Illari, R. Avanzato, C. Napoli, Reducing the psychological burden of isolated oncological patients by means of decision trees, in: *CEUR Workshop Proceedings*, volume 2768, 2020.
- [16] G. Guo, H. Wang, D. Bell, Y. Bi, K. Greer, Knn model-based approach in classification, in: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2003, pp. 986–996.
- [17] K. P. Murphy, et al., Naive bayes classifiers, *University of British Columbia* 18 (2006) 1–8.
- [18] H. Aktaş, N. Çağman, Soft sets and soft groups, *Information sciences* 177 (2007) 2726–2735.
- [19] R. C. De Amorim, B. Mirkin, Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering, *Pattern Recognition* 45 (2012) 1061–1075.
- [20] D. Ayres-de Campos, J. Bernardes, A. Garrido, J. Marques-de Sa, L. Pereira-Leite, Sisporto 2.0: a program for automated analysis of cardiotocograms, *Journal of Maternal-Fetal Medicine* 9 (2000) 311–318.