© 2019 IJSRSET | Volume 6 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology
DOI : https://doi.org/10.32628/IJSRSET196445

Alimentation Status Classification System Using KNN and Naive Bayesian Classifiers

Htwe Htwe Pyone[1], Hnin Yu Yu Win[2]
[1]Faculty of Computer Science, University of Computer Studies (Myitkyina), Myanmar
[2]Faculty of Computer Science, University Computer Studies (Taungoo), Myanmar

ABSTRACT

Nowadays, there is an increasing interest about child health care in the developing countries. Because the future of each country is based on many youths, they need to be healthy. So, it is needed to takecare the health of each child since childhood. Good alimentation is prime important in the attainment of normal growth and development, and in the maintenance of health throughout life. Especially in their earlier live, the children need adequate quantity and appropriate quality of food to meet the alimentation requirement for their physical, mental growth and development. So, this system is proposed as the alimentation status classification system for children. This system classifies the child who has alimentation status or mal-alimentation by using k-nearest neighbor (KNN) and naive bayesian classifiers. Moreover, this system compares the performance of these two classifiers to know which classifier is more precise than other for child's alimentation status classification.

Keywords: Alimentation, K-Nearest Neighbor, Naive Bayesian, Classification

I. INTRODUCTION

In early childhood, adequate alimentation is essential to ensure healthy growth, a strong immune system, and neurological development. Human development and economic growth require well-nourished populations who can learn new skills, think critically and contribute to their communities. Child mal-alimentation impacts cognitive function and contributes to poverty through impeding individuals' ability to lead productive lives. So, alimentation status classification system is important for child. In medical applications, data mining is popular and useful technique.

Data mining is the computer-assisted process of digging through and analysing enormous sets of data and then extracting the meaning of data. By applying data mining methods for medical domains, it can minimize the clinical cost. Computer based medical system is playing a role in assisting health care service. When the computer based medical system uses to process data, it needs to choose what data is useful and what kind of new information which can get from the database. Data mining method is to help and guide in order to get quick and correct instructions on symptoms. This mining method is also to detect data analysis that can be used to extract models describing import data class or to predict future data trends.

Classification is one of data mining methods which are used to predict the class label of unknown instant. Classification approaches consists of decision tree, decision rules, k-nearest neighbor (KNN), naive bayesian, neural network, regression-based method and vector-based method.

Among them, this system uses the KNN and naive bayesian classifiers for alimentation or mal-alimentation status classification. As a result, the proposed system is imitated as a doctor for every family. This system can support to track child growth and to establish appropriate alimentation care plans. This system also supports the junior doctors when they classify the alimentation status problem for children.

II. RELATED WORK

In 2015, S. Indhumathi and G. Vijaybaskar [1] described the intelligent heart disease prediction system by using data mining modeling technique, naive bayes classifier. It is implemented as the web based questionnaire application. Based on the user answers, it extracted hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It answered the complex queries for diagnosing heart disease and thus assisted healthcare practitioners. By providing effective treatments, it helped to reduce treatment costs.

In 2017, H. Bohra, A. Arora and P. Gaikwad [2] provided user an interface for medical guidance. Platform independent system helped people to use it on their computers and provided users an instant guidance on their health issues. This system used naive bayes algorithm that predicted the disease and the daily hygiene diet. By using this system, users will be able to contact the specialist doctors nearby. This system helped users for easy medical treatment and diagnosis.

In 2018, P. A. Sarkate and A. V. Deorankar [3] presented chemical medicine or drug classification by using k-nearest neighbor (KNN) classifier. In the medical domain, classification involves multi classes, heterogeneous and complex data structure. KNN classifier is the most popular, very simple, highly effective algorithm. This classifier is used to solve classification and optimization problem about drug data set categorization.

III. CLASSIFICATION

One of the data mining techniques is classification. Classification is used to analysis the data sets and classify the data sets by the instance and separate the classes according to the instance of data sets. Classification also used to absorb the models which are important data in the data set. Classification has two types of process. First process is the model from data set or trained data set, developed by the applying of classification algorithm. Second process is the model is verified to the predefined model to get accurate and perfect data set or test set. Training set contains set of attributes. Classification process is to find the model from the trained data set. Classification algorithm produces the relationship between two attribute. The main goal of classification algorithm is to maximize the accuracy ratio of the data set [4]. Classification process is shown in Figure 1.

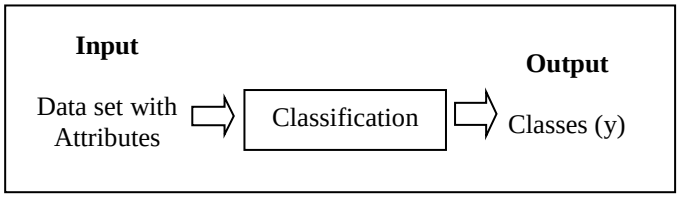

Figure 1: Classification Process [5]

Classification process is a model finding process that is used for portioning the data into different classes according to some constrains. Several major kinds of classification algorithms include decision tree, k-nearest neighbor, support vector machine and naïve bayesian classifiers [6].

IV. K-NEAREST NEIGHBOR (KNN) CLASSIFIER

K-nearest neighbor (KNN) classifiers are based on learning by analogy. Training samples are described by n dimensional numeric attributes. Each sample represents a point in an n-dimensional space. All

training samples are stored in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance, where the Euclidean distance, where the Euclidean distance between two points, $X=(x_1, x_2,....., x_n)$ and $Y=(y_1, y_2,...., y_n)$ is

$$d(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

KNN classifier algorithm is as follows:

- Step 1: determine k
- Step 2: calculate the distance between the new input and all the training data
- Step 3: sort distance and determine k nearest neighbors based on k^{th} minimum distance (Euclidean distance)
- Step 4: gather the categories of those neighbors
- Step 5: determine the categories based on majority vote

KNN classifier assigns the unknown sample with the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space [7].

V. NAIVE BAYESIAN CLASSIFIER

Naive bayesian classifier is a probabilistic classifier, which considers probability values to classify attributes. It is independent feature model. It calculates probabilities to hypothesis. And also, it is robust to noise in data. This classifier assumes that presence/ absence of a particular feature is lineal to any feature's presence/ absence. It is a supervised learning classifier. It is easy to construct and easy to interpret. So, users who are don't have prior knowledge in classification can also construct naive bayesian classifier [8].

Processing steps of naive bayesian are as follows:

1. Each data sample is represented by n-dimensional feature vector, $X=(x_1, x_2 \ldots x_n)$ depicting n-measurements made on the sample from n-attributes, respectively, $A_1, A_2, \ldots, A_n$.\

2. Suppose that there are m classes, $C_1, C_2, \ldots, C_m$. Given an unknown data sample X, Naïve Bayesian classifier assigns an unknown sample X to the class C_i if and only if

$$P(C_i \backslash X) > P(C_j \backslash X) \text{ for } 1 \le j \le m, j \ne i \quad (2)$$

The class C_i for which $P(C_i \backslash X)$ that is maximized, called maximum posteriori hypothesis. By Bayes theorem,

$$P(C_i \backslash X) = P(X \backslash C_i)\, P(C_i)/ P(X) \quad (3)$$

3. As P(X) is constant for all classes, only $P(X \backslash C_i)\, P(C_i)$ need to be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, $P(C_1) = P(C_2) = \ldots\ldots = P(C_m)$ and we would therefore maximize $P(X \backslash C_i)\, P(C_i)$.
4. Given data sets with many attributes, it would be extremely expensive to compute $P(X \backslash C_i)$. In order to reduce computation in evaluating $P(X \backslash C_i)$, the Naïve assumption of class conditional independence is made. Thus,

$$P(X \backslash C_i) = \prod_{k=1}^{n} P(x_k \backslash C_i) \quad (4)$$

The probability $P(x_1 \backslash C_i), P(x_2 \backslash C_i), \ldots P(x_n \backslash C_i)$ can be estimated from the data samples.

5. In order to classify an unknown sample X, $P(X \backslash C_i)\, P(C_i)$ is evaluated for each class C_i,. Sample X is then assigned to the class C_i if and only if

$$P(X \backslash C_i)\, P(C_i) > P(X \backslash C_j)\, P(C_j) \quad \text{for } 1 \le j \le m, j \ne i \quad (5)$$

In other words, it is assigned to the class C_i for which $P(X \backslash C_i)\, P(C_i)$ is the maximum [9].

VI.PROPOSED SYSTEM DESIGN

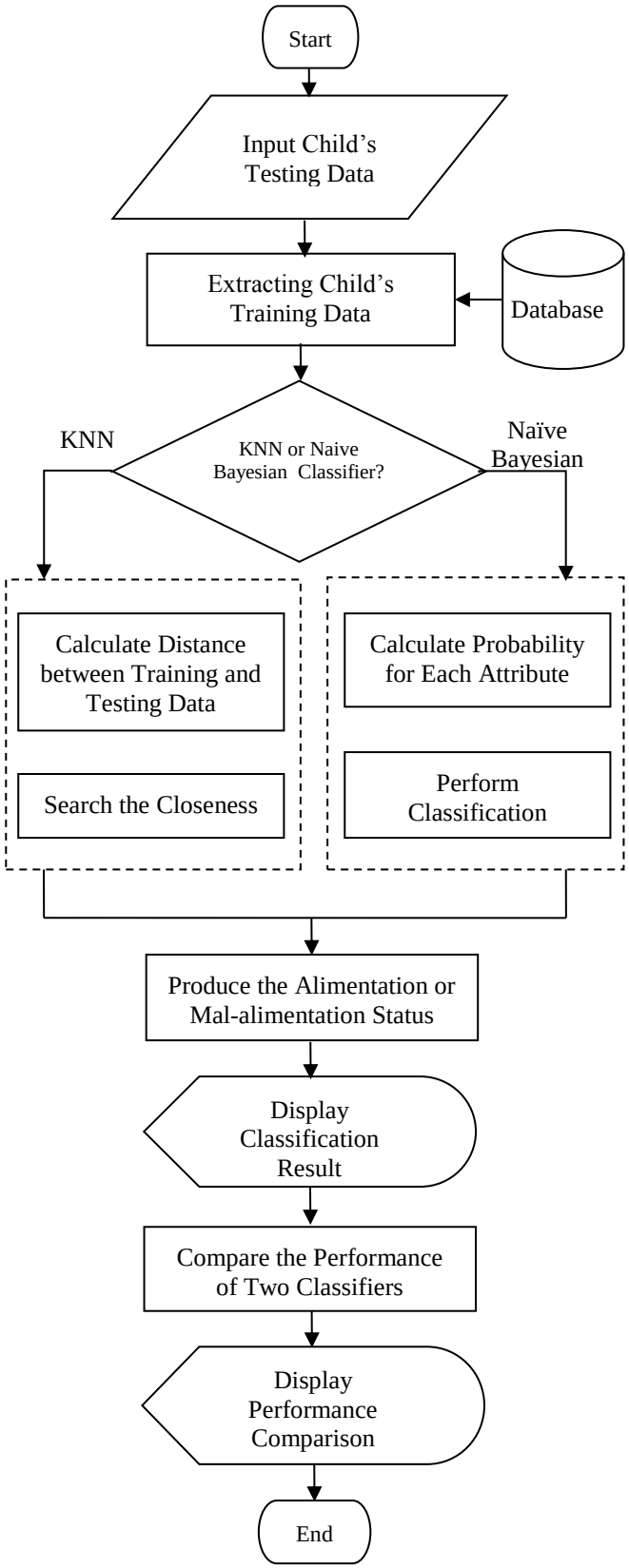

Figure 2 : Proposed System Design

This system is proposed as the alimentation status classification system by using k-nearest neighbor (KNN) and naive bayesian classifiers. Firstly, the user inputs the child's testing data for alimentation classification. After receiving the inputted data, this system extracts the child's training data from the database. By using these training and testing data, this system performs the classification.

In the KNN classification process, this system calculates the distance between testing data and each training data by using Euclidean distance method. Based on the distance results, this system searches the closeness for classification. In the naive bayesian classifier, this system calculates the probability for each attribute. Using these probability results, this system performs the classification process. After finishing each classification process, this system compares the performance of each classifier to know which classifier is more precise than other.

A. Explanation of the System

As a sample, this system tested 10 child's training record with 17 features. These child's alimentation features are shown in Table 1.

TABLE I

CHILD'S ALIMENTATION FEATURES

ID	Attribute Name	Attribute Properties
1	Sex	Male, Female
2	Age	6 months – 5 years
3	Weight (lb)	15-50
4	Height (cm)	35-110
5	Mother Education	Illiterate, Read and Write, Primary school passed, Middle school passed, High school passed, University level and Graduate
6	Family Numbers	2-20
7	Sibling	0-10
8	Monthly Family Income	Low (<100000kyats), Middle (100000-300000kyats), High (>300000kyats)

9	Extra Food per Day	Snack and fast food/Other (One/Two/Three Times), None
10	Diarrheal disease	Yes, No
11	Illness	Yes, No
12	Immunization	Complete, Incomplete and Not known
13	Alimentation Knowledge	Yes, No
14	Breast Feeding	Yes, No
15	Feed per Day (Rice and Curry)	Two/ Three/ Four Times
16	Consultant	Specialists, GP doctors, Basic health staff, Traditional, Medicine and Quack
17	Class	Alimentation, Mal-alimentation

Sample child's alimentation training dataset with 10 records are shown in Table 2.

TABLE III

CHILD'S ALIMENTATION TRAINING DATASET

ID	Sex	Age	Weight	Height	..	Class
1	Male	3-4 year	>16kg	90.1-100cm	..	Alimentation
2	Female	4-5 year	13.1-16kg	> 110 cm	..	Alimentation
3	Female	3-4 year	<=10kg	80.1-90cm	..	Mal-Alimentation
4	Male	4-5 year	13.1-16kg	80.1-90 cm	..	Alimentation
5	Male	3-4 year	>16kg	90.1-100cm	..	Mal-Alimentation
6	Female	3-4 year	13.1-16kg	80.1-90 cm	..	Alimentation
7	Male	4-5 year	<=10kg	80.1-90 cm	..	Mal-Alimentation
8	Male	3-4 year	13.1-16kg	> 110 cm	..	Mal-Alimentation
9	Female	4-5 year	13.1-16kg	> 110 cm	..	Alimentation
10	Female	4-5 year	>16kg	90.1-100cm	..	Alimentation

The user inputs the child's testing data that includes sex is "Female", age is "4-5 year", weight "13.1-16kg", height is "90.1-100cm", mother education is "Graduate", family number is ">=7", family income is "Middle", sibling is ">=7", food is "Rice and curry", diarrheal disease is "No", illness "Yes", immunization is "Not known", nutritional knowledge is "No", breast feeding is "No", feed per day "Three times" and consultant is "traditional". Then, this system classifies this testing data by using KNN and naive bayesian classifiers.

The results of naive bayesian classifier are as follows:

- P(Class = Alimentation) = 6/10 = 0.6
- P(Class = Mal-alimentation) = 4/10 = 0.4
- P(X|Class = Alimentation) = 0.667 × 0.667 × 0.667 × 0.333 × 0.167 × 0.5 × 0.167 × 0.333 × 1 × 1 × 0.167 × 0.167 × 0.167 × 0.167 × 0.5 × 0.167 = 0.0000000298
- P(X|Class = Mal-alimentation) = 0.25 × 0.25 × 0.25 × 0.25 × 0.5 × 0.75 × 0.75 × 0.5 × 0.25 × 0.25 × 0.5 × 0.5 × 0.75 × 0.75 × 0.25 × 0.25 = 0.0000003017
- P(X|Class = Alimentation) P(Class = Alimentation) = 0.0000000298 × 0.6= 0.00000001788
- P(X|Class = Mal-alimentation) P(Class = Mal-alimentation) = 0.0000003017 × 0.4= 0.00000012068

The results of KNN classifier are as follows:

- Distance (child testing and training record 1) = 1
- Distance (child testing and training record 2)= 1.5
- Distance (child testing and training record 3) = 2.828

- Distance (child testing and training record 4) = 1
- Distance (child testing and training record 5) = 2.236
- Distance (child testing and training record 6) = 1
- Distance (child testing and training record 7) = 1.732
- Distance (child testing and training record 8) = 3
- Distance (child testing and training record 9) = 1.3
- Distance (child testing and training record 10) = 1

After finishing each classifier process, this system produces the "Mal-Alimentation" result for user inputted testing data.

B. Experimental Results of the System

For performance measurement, this system is tested 100 testing data. To access the "accuracy" or "correctness" of the system, this system uses the following equation.

$$A_i = t/n * 100 \quad (6)$$

A_i is the accuracy of the system. The "t" is the number of corrected testing data. The "n" is the number of testing data.

In the KNN classification process, this system obtains the correct rate that is 93% and the error rate that is 7% respectively. In the naïve bayesian classification process, this system obtains the correct rate that is 88% and the error rate that is 12% respectively. According to the experimental result, the performance of KNN classifier is more precise than naïve bayesian classifier.

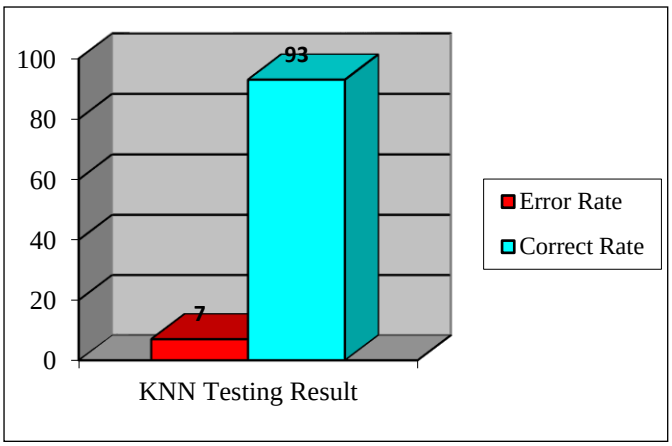

Figure 3 : KNN Testing Result

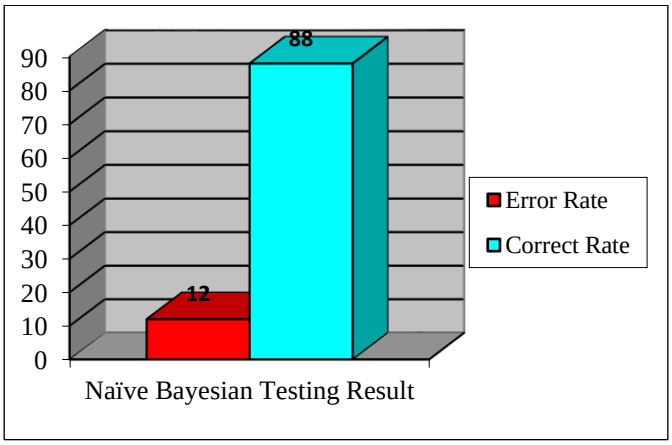

Figure 4 : Naïve Bayesian Testing Result

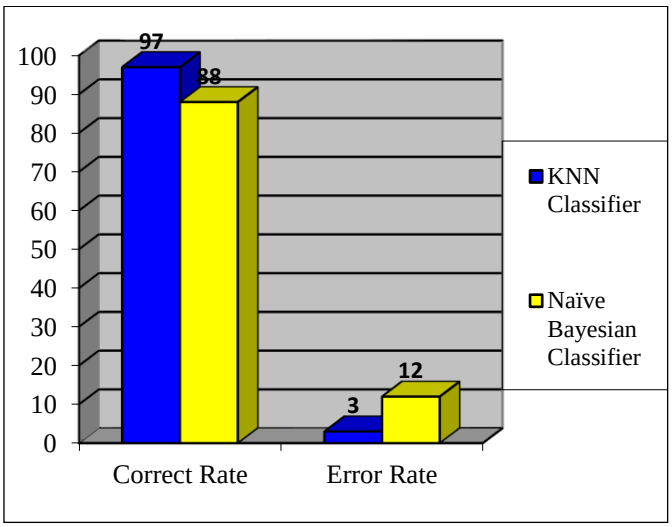

Figure 5 : Performance Comparison Result

VII. CONCLUSION

Classification system for child' alimentation status ages of between 6 months and 5 years is implemented by using k-nearest neighbor and naive bayesian classifiers. The system can classify the alimentation status of the children quickly and correctly. It helps parents for testing their children's alimentation and mal-alimentation status. And, the parents can know about the child's health knowledge. Then, it also supports the junior doctors when they classify the alimentation status problem. Besides, it can save time and money because parents don't need to go and consult the child specialists.

VIII. REFERENCES

[1] S. Indhumathi and G. Vijaybaskar, "Web Based Health Care Detection using Naive Bayes Algorithm", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 4, no. 9, 2015.

[2] H. Bohra, A. Arora and P. Gaikwad, "Health Prediction and Medical Diagnosis using Naive Bayes", International Journal of Advanced Research in Computer and Communication Engineering, vol. 6, no. 4, 2017.

[3] P. A. Sarkate and A. V. Deorankar, "Classification of Chemical Medicine or Drug using K Nearest Neighbor (KNN) and Genetic Algorithm", International Research Journal of Engineering and Technology (IRJET), vol. 5, no. 3, 2018.

[4] A. Lourducaroline, S. Manikandan and D. Kanniamma, "Comparative Study of Classification Algorithms for Data Mining", International Journal of Engineering Science Invention (IJESI), 2018.

[5] S. Umadevi and K. S. Jeen Marseline, "A Survey on Data Mining Classification Algorithms", International Conference on Signal Processing and Communication (ICSPC' 17), IEEE, 2017.

[6] R. Kumar, "Classification Algorithms for Data Mining: A Survey", International Journal of Innovations in Engineering and Technology (IJIET), vol. 1, no. 2, pp. 7-14, 2012.

[7] N. P. Thair, "Survey of Classifiction Techniques in Data Mining", Proceedings of the International Multiconference of Engineers and Computer Scientists (IMECS), 2009.

[8] K. P. Jyothi, R. Sivaranjani and T. K. Mishra, "A Study of Classification Techniques of Data Mining Techniques in Health Related Research", International Journal of Innovative Research in Computer and Communication Engineering, vol. 5, no. 7, 2017.

[9] Jiawei Han, Micheline Kanber and Jian Pei, "Data Mining Concepts and Techniques (Third Edition)", 2001.

Cite this article as :

Htwe Htwe Pyone, Hnin Yu Yu Win, "Alimentation Status Classification System Using KNN and Naive Bayesian Classifiers", International Journal of Scientific Research in Science, Engineering and Technology (IJSRSET), Online ISSN : 2394-4099, Print ISSN : 2395-1990, Volume 6 Issue 4, pp. 346-352, July-August 2019. Available at doi : https://doi.org/10.32628/IJSRSET196445

Journal URL : http://ijsrset.com/IJSRSET196445