

KNN Algorithm for Consulting Behavioral Disorders in Children

Mohammad Bazmara¹, Sania Vahedian Movahed², Samira Ramadhani³

^{1,2}School of Electrical and Computer Engineering Shiraz University, Shiraz, Iran

³Department of Computer Engineering and Information Technology, Isfahan University of Technology, Isfahan, Iran

Received: September 27 2013

Accepted: November 13 2013

ABSTRACT

In this paper the KNN algorithm is used to detect behavioral disorders in children. In fact users can use it in consultation services in the field of children behavioral disorders. The result of consultation can be observed after responding to the questions. Finding important features by an expert is one of the most difficult steps in creating this algorithm. Then psychiatric information of children who have been referred to a special clinic for consultation and treatment is gathered and the model has been implemented. Using a very simple user interface, a psychology expert can create the sequence of questions and responses, to diagnose and consult the behavioral disorders in children. The model transfers the decision to the patients without the psychiatrist's intervention. Experimental results indicate that the proposed model has the necessary efficiency and accuracy to counsel the patient in the absence of the psychiatrist.

KEYWORDS: Child Psychology, Nearest Neighbor Algorithm, Child Consultation, Behavioral Disorders

I. INTRODUCTION

Nowadays, consulting software services is one of the most attractive and popular software services all over the world. Such services are presented on various issues like marriage and social issues, rights and civil issues, investment and stock purchase, psychology and health and other various topics. Depending on the related subject, each of software can be provided by its own procedures and methods. Various Methods and ways of presenting these services can be classified in to the following three groups. However it should be noted that this classification may not cover all available cases and other methods may exist too. This classification has been done based on the current methods in available Persian applications and websites. Usually a combination of these methods is used by software and websites to present their services [1, 2].

A. First method: presenting articles and literature related to the subject

Some software and websites collect and classify articles and related recommendations, to provide the possibility of searching and finding results for their users. Moreover, some websites create a Frequently Asked Questions (FAQ) of common user questions. In FAQ, some predetermined questions are answered, which is related to the intended topic. Although this method cannot be considered as one of the categories of consulting services, it was also mentioned in this article because users can find answers to some of their questions by using this method.

B. Second method: Using Chat Rooms and Forums

Chat rooms are a virtual site in consulting websites, providing the possibility of online textual, audio or video interlocution among users and counselors. In this room, users ask their questions and the counselor offers his ideas. While in forums, each room has specified topic for discussion users must raise their questions, then an expert or consuler non-simultaneously investigates the topic, if it is needed to ask some questions from users, then the final answer and consultation are announced. This model of information exchange is also observed in some weblogs with a specialized approach [3].

However in this method, expert directly or indirectly studies the topic and provides the result, but the process of offering consultation has several fundamental problems. At first, it is a time-consuming process. Since each expert answers to the users within a limited time, hence the responding time increases with respect to the number of users. Increasing the response time is in contrast with the principle of removing time limitations in e-commercial services. Second problem is the textual nature of questions and answers that there is no method to classify and analyze the information and consultation results. Moreover in such sites before understanding the way of raised issues and getting results, users have to pay consoling fee, which is inconsistent with the principle of enhancing customer satisfaction in e-commerce.

C. Third method: Standard Questionnaires

In this method, standard predetermined questionnaires are given to the users. The result of consultation is determined with respect to their answers. Although more software techniques have been used compare to the

*Corresponding Author: Mohammad Bazmara, School of Electrical and Computer Engineering, Shiraz University.
E-mail: Mohamad.bazmara@gmail.com

second method, but also has some problems in the performance of software and active sites in this field. At first provided questionnaires are usually the same for all users and there is no category like the age or sex of users. Using predetermined multiple-answers not including all possible alternatives is another problem. The last and most important problem is that the relation between available disorders was not considered. Despite of any mental and physiological complications can interact with each other, in this method each disorder investigates individually and the effect of interactions is not considered.

Is it feasible to create a psychological software program so that it can individually examine each user like an expert and present the result of the examination after receiving questions? In fact, in this paper, we tried to find the answer to this question. So a research project was proposed to create software for consultation of behavioral disorders in children. Children behavioral disorders were chosen in this paper for two reasons. Firstly, some of psychology experts believe in therapy skills and they claim the majority of behavioral disorders in children are due to the deficiency or weakness in one or few behavioral skills. These disorders are not caused by physical or mental complications. Therefore, it is not necessary to prescribe any drug. Such the problems can be removed only by notifying parents and teaching how to transfer a skill to children. Secondly this group of psychologists does not require any physical examination of the child and often speak with children's families about their behaviors to find their skill shortages. Thus, instead of presence examination this process is virtually possible with the least possible risk.

The rest of paper is organized as follows: related works are described in section II. In section III the motivation of this paper is shown. Section IV briefly presents KNN. The application of KNN in consulting behavioral disorders in children is described in section V. Section VI includes methodology and evaluation and results are given in section VII. Paper is concluded in section VIII.

II. RELATED WORK

After determining the issue, solutions should be looked for. We know that using modern technologies and innovation in commercial models, the pioneer companies could be developed and improved themselves. Undoubtedly, one of the fundamental changes in IT is researches' achievements in artificial intelligence. AI provides various methods to be applied in different areas such as really finding soccer talent and goalkeeper quality [4, 5]. Its applications in bioinformatics [6, 7], agriculture [8, 9] or selecting systems of important features [10] and also geographic information systems using fuzzy uncertainty [11].

III. MOTIVATION

After holding several meetings with a psychology expert and discussing his needs for better implementation of the project, it turned out that the nearest neighbor algorithm which is a sub-branch of artificial intelligence could be used for implementation. Creating a software program based on the nearest neighbor algorithm has its own specific problems. One of the major problems is to determine the value of K , one of the input parameters of algorithm. The other important part is to determine what kind of distance measurement must be used. The concept of closeness can be varied depending on the type of distance measurement. Some interior steps like doing consecutive discussion meetings, filling in various forums, and several other methods to transfer knowledge from an expert to program designer, make data collection, caused that implementation and finally testing the software become very complex [12]. Hence, in this paper, we are looking for an intelligent software environment so that a comfortable environment with adequate facilities can be provided for the expert to create an intelligent system. Generally, there are two main working groups in the software including users group and experts group. Levels of accessibility and workflow of each of them are studied in the following.

IV. KNN

This algorithm assumes the n -dimensional space R^n for all instances. Standard Euclidean distance is used to define the nearest neighbors of an instance. More precisely, let x be an arbitrary instance described by the feature vector

$$\langle a_1(x), a_2(x), \dots, a_n(x) \rangle \quad (1)$$

Where $a_r(x)$ denotes the value of the r th attribute of instance x . Then the distance between two instances x_i and x_j is defined to be $d(x_i, x_j)$ where

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2} \quad (2)$$

In nearest-neighbor learning the target function may be either discrete-valued or real-valued. Consider learning discrete-valued target functions of the form $f: R^n \rightarrow V$, where V is the finite set, $\{v_1, v_2, \dots, v_s\}$, the k -Nearest Neighbor algorithm for approximating a discrete-valued target function is given in figure 1. The value $\tilde{f}(x)$ returned by this algorithm as its estimate of $f(x)$ is just the most common value of f among the k training

examples nearest to x . If we choose $k = 1$, then the 1-Nearest Neighbor algorithm assigns to $f(x)$ the value $f(x_i)$ where x_i is the training instance nearest to x . For larger values of k , the algorithm assigns the most common value among the k nearest training examples [14].

Training algorithm:

- For each training example $(x, f(x))$, add the example to the list *training_examples*

Classification algorithm:

- Given a query instance x_q to be classified,
 - Let $x_1 \dots x_k$ denote the k instances from *training_examples* that are nearest to x_q
 - Return

$$\hat{f}(x_q) \leftarrow \operatorname{argmax}_{v \in V} \sum_{i=1}^k \delta(v, f(x_i))$$

where $\delta(a, b) = 1$ if $a = b$ and where $\delta(a, b) = 0$ otherwise.

Figure 1: The k-Nearest Neighbor algorithm for approximating a discrete-valued function $f: \mathbf{R}^n \rightarrow \mathbf{V}$.

V. THE APPLICATION OF KNN IN CONSULTING BEHAVIORAL DISORDERS IN CHILDREN

When trying to solve new problems, individuals often refer to solutions of the same issues previously resolved. Nearest neighbor algorithm is a version of classification technique. In this decision-making method allocating a group into a new case is done by studying some numbers of (k) from the most similar cases or neighbors. The number of cases is counted for each class and a new case is allocated to the group that a larger number of neighbors belong to it [13-15]. For example, there are three classes, w_1 , w_2 and w_3 of data in the figure 2.

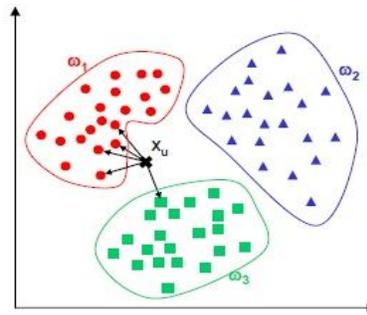


Figure 2: An example of an unknown data allocation with the nearest neighbor algorithm

Assume that each data represents a child. The green group (w_3) belongs to a class of children with no hyperactivity, red group (w_1) belongs to a class with low hyperactivity and blue group (w_2) belongs to a class with extreme hyperactivity. If k is considered 5 (this amount can vary in terms of expert's view) and we want to find the Behavioral Disorders of child X, 5 inputs (children) being near (having more similarity) to our input (our desired children) are considered. It is obvious that 4 closer inputs belong to w_1 class or the class without mental disorder and only 1 input belongs to w_3 class or with no hyperactivity. Then it is concluded that our desired child has no behavioral disorder. To better understand, the flowchart of K nearest neighbor algorithm is shown in Figure 3.

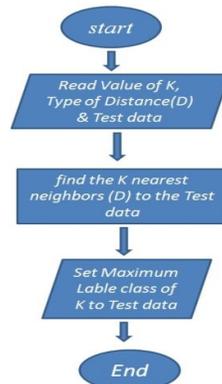


Figure 3: The flowchart of K nearest neighbor classifier procedure[2].

After determining k , it's necessary to specify what type of distance measure could be used. In this paper various distance measures are examined like Euclidean distance which is one of the most famous measures. Results have been reported in this paper.

VI. METHODOLOGY

First, the required inputs and questions should be found to know the rate of behavioral disorder in children. The Questions were considered as the criteria that the psychologist asks face to face from children's parents for the diagnosis. The operation was performed step by step to reach the desired result. In the process of consultation, the main problem is to determine which questions the user should respond to. So these questions have been considered as the input features. After displaying questions to the user, the system should wait for responses to make a consultation.

In this application, according to the psychology expert request, hyperactivity was considered among from three very common disorders- attention-deficit, hyperactivity and impulsivity. The expert considered a total of 9 questions to determine the disorder which is shown in table 1.

Table 1: List of questions related to hyperactivity

The Name of the Question	Question
Wiggling	Does he sit quietly and control the movement of his hands and legs?
Observing regulations	Does he sit when it is socially or based on the class rules required?
Regulating the movement	Does he regulate his movement activity?
Slow playing	Does he play slowly?
Follow up	Does he do his work seriously and pursue it?
Regulating word	Does he regulate his verbal activity?
Contemplating	Does he contemplate on questions?
Turn taking	Does he wait for his turn?
Interrupting and interfering	Does he participate in play or conversation without interrupting other conversation or interfering with others' playing?

In psychological resources, the answers for these questions are considered as yes or no, while it is seen that families often cannot certainly give yes or no answers. So we asked them to allocate a number in the range of [0-100] to each question.

This diagnosis method has high efficiency. After receiving inputs, the nearest individuals to the child previously diagnosed his behavioral disorder finds and depending on their diagnosis, a new decision is made for a new child.

VII. EVALUATION & RESULT

For consulting the behavioral disorders in children, the system of nearest neighbor algorithm was performed in two Psychology Clinics in Tehran then studied by experts selected among from specialists and university professors.

A trial group consisted of 120 children with the age range of 3-7 years were selected. According to explanations of his parents and the psychologist, each child received a number for every 9 questions which one state of without hyperactivity, low hyperactivity and extreme hyperactivity was diagnosed for them.

Now, the classification of the nearest neighbor should be performed on them. In so doing, two factors should be considered: one was the value of k and the other was the selection of the desired distance measure. Since there was no equation for determining these two factors in the classification of the nearest neighbor, various states were considered so that we could know which distance was more appropriate with which K .

K was obtained from the most common case used in papers. Also, for the distance, the following four known distances were used:

- 1) Euclidean: Euclidean distance
- 2) Cityblock: the sum of absolute values for distances
- 3) Cosine: one subtracting the cosine of angle enclosed between two points (a point is considered as vector)
- 4) Correlation: one subtracting correlation between two points

That is defined as follows:

- **Euclidean distance:** The *Euclidean distance* between two points, a and b , with k dimensions is calculated as:

$$\sqrt{\sum_{j=1}^k (a_j - b_j)^2} \quad (3)$$

- **Cityblock distance:** The *City block distance* between two points, a and b , with k dimensions is calculated as:

$$\sum_{j=1}^k |a_j - b_j| \quad (4)$$

- **Cosine distance:** The *Cosine distance* between two points, a and b , with k dimensions is calculated as:

$$\frac{\sum_{j=1}^k a_j \times b_j}{\text{norm}(a) \times \text{norm}(b)} \tag{5}$$

Where

$$\text{norm}(a) = \sqrt{\sum_{j=1}^k a_j^2} \quad \& \quad \text{norm}(b) = \sqrt{\sum_{j=1}^k b_j^2}$$

- **Correlation distance:** The *Correlation distance* between two points, *a* and *b*, with *k* dimensions is calculated as:

$$\frac{\text{cov}(a,b)}{\text{std}(a) \times \text{std}(b)} \tag{6}$$

where

$$\text{cov}(a,b) = \frac{1}{k} \sum_{j=1}^k (a_j - \bar{a}) \times (b_j - \bar{b})$$

$$\bar{a} = \frac{1}{k} \sum_{j=1}^k a_j \quad \& \quad \bar{b} = \frac{1}{k} \sum_{j=1}^k b_j$$

To obtain the diagnosis accuracy percentage for the algorithm, each time one input was removed from the input set and considered as the test input. Then, according to *k* and the desired distance, the algorithm attributed it to a class. If the allocated class was similar to the psychiatric diagnosis, we would say that the diagnosis is correct; otherwise it would be said that the diagnosis is wrong.

The results are provided in table 2.

Table 2: The percentage of correct diagnosis of players' post with classification of the nearest neighbor

Distance \ K	K=1	K=3	K=5	K=10	AVG
Euclidean	73	79	79	83	78.5
Cityblock	72	79	82	79	78
Cosine	87	90	92	93	90.5
Correlation	83	89	92	84	87
AVG	78.75	84.25	86.25	84.75	

For better understanding of the accuracy of data, the table is presented visually in figure 4.

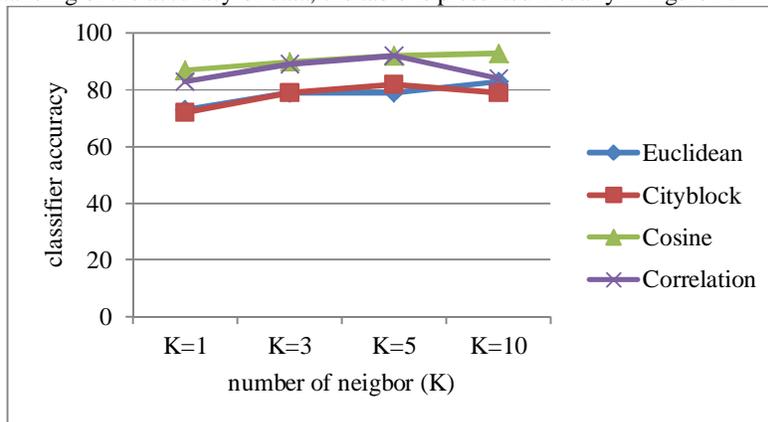


Figure 4: the nearest neighbor classifier accuracy according to different values of *K* parameter and distance type.

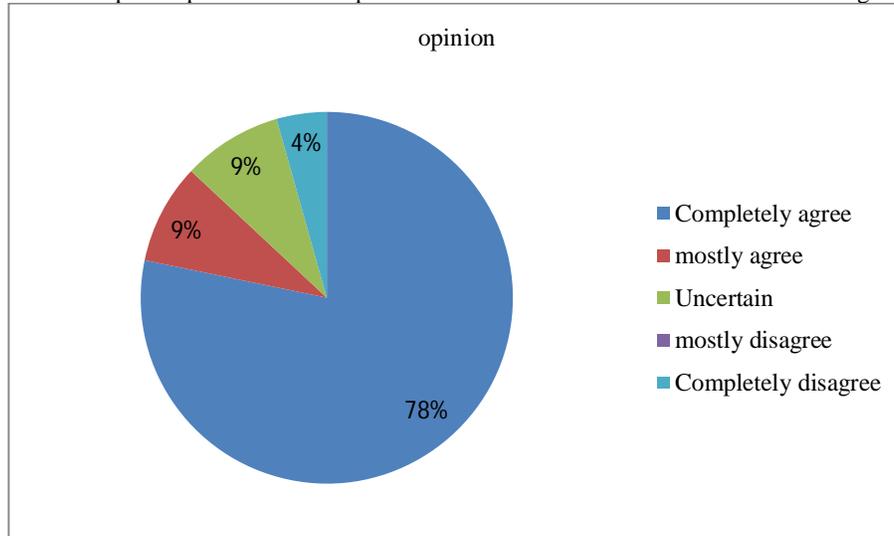
From results in table 2 and figure 4, the followings can be concluded:

- It is better to consider *k* as 5 or 10
- It is better to use distances of cosine for the classification and then use correlation as the next choice

23 experts with the best ranking were asked to describe outputs related to the classification of the nearest neighbor. They can have five possible descriptions:

Completely agree, mostly agree, uncertain, mostly disagree or completely disagree. Results are presented in Table3.

Table 3: Experts' opinion about the performance of the classification of the closest neighbor



VIII. CONCLUSION

The KNN algorithm was used in the diagnosis of the behavioral disorders in children in that users can refer to for benefiting from consultation services in the field of the behavioral disorders in children and observe the consultation result after responding to the provided questions. Then, the model was implemented in which a psychology expert can create the sequence of question and responses required for diagnosis and consultation of behavioral disorders in children using a very simple user interface and also, the transference his decision-making pattern to the user (patient) without psychiatrist's intervention. Finally, the obtained results indicate that the accuracy of the method was acceptable. It concluded that by using $k=5$ and correlation as distance measure; the system had the highest accuracy. The Table 2 shown that accuracy of such system was 92%. Also, experts in psychology confirm this method.

IX. FUTURE WORKS

To have more powerful system than before, online services can be used. So users can access to this software more easily. Also other powerful methods of artificial intelligence like fuzzy logic, neural network and evolutionary algorithms can be used to make this software more useful. Also system can be designed b interpretability to express the reasons of choosing each answer.

Acknowledgment

The authors declare that they have no conflicts of interest in this research.

REFERENCES

1. C. Ghezzi, M. Jazayeri, and D. Mandrioli, *Fundamentals of software engineering*, Prentice Hall PTR, 2002.
2. Carmona, Cristóbal J., et al. Web usage mining to improve the design of an e-commerce website: OrOliveSur. com, *Expert Systems with Applications* 2012: 11243-11249.
3. W. Zhang, X.J. Tang, and T. Yoshida, "Web text mining on a scientific forum," *International Journal of Knowledge and Systems Sciences* 2006, pp. 51-59.
4. M. Bazmara, and S. Jafari, "K Nearest Neighbor Algorithm for Finding Soccer Talent," *Journal of Basic and Applied Scientific Research* 2013, pp. 981-986.
5. M. Bazmara, and S. Jafari, "A Fuzzy expert system for goalkeeper quality recognition," *International Journal of Computer Science Issues* 2012, pp. 318-322.
6. Y. Huang, and Y. Li, "Prediction of protein subcellular locations using fuzzy k-NN method," *Bioinformatics* 2004, pp. 21-28.

7. S. Ramadhani, S.R. Mousavi, and M. Talebi, "An improved heuristic for haplotype inference," *Gene*2012, pp. 177-182.
8. H. Azadi, J. van den Berg, M. Shahvali, and G. Hosseininia, "Sustainable rangeland management using fuzzy logic: A case study in Southwest Iran," *Agriculture, Ecosystems & Environment*2009, pp. 193-200.
9. H. Monga, "Classification of Biological Species Based on Leaf Architecture-A review," *International Journal of Computer Science and Information Technology & Security*2012, pp. 332-334.
10. I. Jamali, M. Bazmara, and S. Jafari, "Feature Selection in Imbalance data sets," *International Journal of Computer Science Issues*2012, pp. 42-45.
11. F. Mohammadi, and M. Bazmara, "A New Approach of Fuzzy Theory with Uncertainties in Geographic Information Systems," *International Journal of Mechatronics, Electrical and Computer Technology*2013, pp. 1001-1014.
12. A. Kaufmann, and M.M. Gupta, *Fuzzy mathematical models in engineering and management science*, Elsevier Science Inc., 1988.
13. Dong, Guozhu, James Bailey, and Jinyan Li, eds. *Contrast Data Mining: Concepts, Algorithms, and Applications*. Vol. 28. CRC Press, 2012.
14. T.M. Mitchell, *Machine learning*. WCB, McGraw-Hill Boston, MA., 1997.
15. T.M. Mitchell, "Machine learning and data mining," *Communications of the ACM*1999, pp. 30-36.