K-Nearest Neighbor Based Association Data Mining in Healthcare Data Systems

무스타파 파워드 (경성대학교 경영학과 박사과정, 주저자 hammadmustafa63@gmail.com)
김종호 (경성대학교 경영학과 교수, 교신저자 jonghokim@ks.ac.kr)

… Abstract …

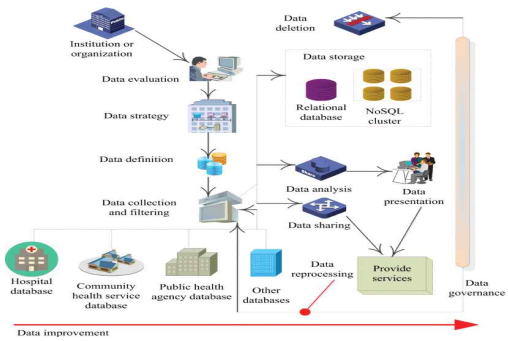
Indeed, data mining techniques have been used to find hidden patterns and linkages in health-related practices, summarize data in unique ways that are both understandable to healthcare stakeholders, and predict future patterns and behaviors. Various data mining strategies and approaches have gotten a lot of attention and research. Medical information has progressed in the direction of intelligence as a result of the rapid advancement of information technology. The K-nearest neighbor classification algorithm is frequently utilized in various disciplines due to its simplicity. When the sample size is high and the feature attributes are substantial, the K- nearest neighbor algorithm classification efficiency has also grown greatly. This study demonstrates how a K-nearest neighbor-based data mining technique was used to index data and analyze a clinical data set from an outpatient facility. As a result, the experimental findings suggest that the proposed technique can significantly increase the KNN algorithm's classification efficiency when processing a huge data set. Using the K-nearest neighbor algorithm, data mining techniques may classify customer behavior depending on the prospect, responder, active, and other entities of the customer's life cycle.

Key Words : Algorithm, Association, Classification, Data Mining, Knowledge, Healthcare, Technique

Ⅰ. Introduction

The enhanced KNN (K-nearest neighbor) method is compared to the traditional KNN method in this study. Each class is given a weight. To ensure that the imposed weight has no detrimental impact on outliers, the approach takes into account the class distribution surrounding the query instance. To overcome the disadvantages of typical KNN algorithms in processing large data sets, this work proposes an upgraded KNN algorithm based on cluster squeezing and density cropping. By boosting the search speed of K-nearest neighbors while maintaining the KNN method's classification accuracy, the approach leverages clustering to accomplish the squeezing process and improves the classification

efficiency of the KNN algorithm. The results of the experiments show that the proposed approach may effectively improve the KNN algorithm's classification efficiency in large data sets while retaining the algorithm's classification accuracy and giving good classification performance. Medical information has advanced in the direction of intelligence as a result of fast advancements in information technology. Big data is a critical data resource for medical service intelligence and smart healthcare in the medical field. For the changes in medical information, classification of medical health big data is critical. Because of its simplicity, the K-nearest neighbor classification algorithm is widely used in a variety of fields. When the sample size is large and the feature characteristics are large, the performance of the KNN technique classification will be significantly improved. User behavior data, which comprises user information, user pages, user transactions, and time-series data from these variables, is instructional data that can be asymmetrical [9]. As shown in figure 1, manual data processing is time consuming and prone to calculation errors. To efficiently handle this data, data mining help is necessary, and a business intelligence system based on customer relationship management can be built for marketing reasons [4]. Data mining techniques can be used to build business intelligence systems that analyze client behavior trends. A vast amount of client data and inconsistencies can be sorted into various classifications using the K-nearest neighbor classification algorithm. The K-nearest neighbor method can also be used to analyze user input faster and provide users with real-time responses and recommendations [5]. The K-nearest neighbor algorithm is one of the most fundamental machine learning approaches. It is based on the simple idea that "items that are 'near' each other will have comparable features." As a result, if you know one of the objects' characteristics, you may predict the attributes of its nearest neighbor. The KNN is a variant of the closest neighbor technique that assumes that any new instance may be categorized by the majority vote of its 'K' neighbors, where k is a positive integer, usually a modest number [15].



<Fig. 1> Scenario of healthcare data extraction

Clinical data is a type of data that is used to hold medical information in the same manner as patient records are used. The amount of data and data sets required to store digital data has increased dramatically. Natural clinical information is frequently extensive and diversified in character. Photographs, patient interviews, research center data, and the doctor's judgements and assessments were all used to compile the information. [13, 28]. Clinical data is now available in a variety of public and private data sets, which was only made possible by new data set improvements and the Internet. It has been estimated that the medical care business produces terabytes of data on a regular basis. In truth, the process of separating useful data for high-quality medical care is fascinating. Despite this, the information extracted from it is almost irrelevant [14].

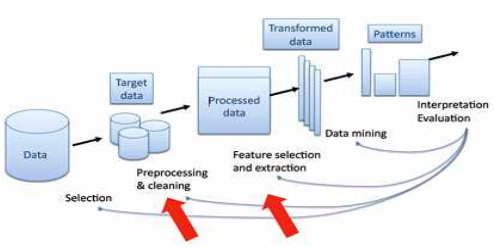
Ⅱ. Literature Review

1. Classification and Association Models

There are numerous data mining models, each of which differs from one application field to the next. In any case, it's more likely to be ordered in two groups. To be more specific, predictive and descriptive models. The association technique searches vast databases for things that are connected or together, and this type of link is known as the association rule [2, 21]. Affiliation rules are commonly used in marketing, where the board, publicizing, and so on are detached from these affiliation rules that exist among different qualities, as illustrated in figure 2. Indeed, affiliation-based information mining goals to uncover links between attributes and then make some rules based on such informational indexes [3].

Furthermore, categorization assigns target classes to data sets. For every patch of the data present, classification methods anticipate the target classes. For example, based on their infection

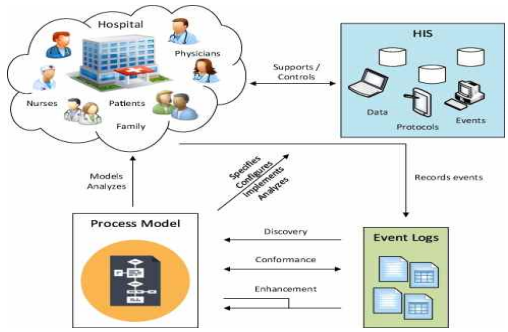
designs, a patient can be classified as "high hazard" or "usually safe" using order approaches. Because the classes are recognized under this system, it is a form of administered learning. Ordering tasks can be done in two ways [1, 26]. During the characterization work, the dataset is divided into two sections: preparing and testing informational collections. In addition, an informational index is used to generate the classifier, and the classifier's correctness is tested on the test dataset. Data mining is a task that is mostly used in the medical care industry [5, 16].



<Fig. 2> Systematic data flow activity

2. Regression Models

Learning a capability to plan an information thing to a genuine esteemed forecast variable is known as regression. Regression, in actuality, create a link among unknown and undefined variables and known dependent variables [10]. To deal with these massive data bases of clinical information and extract useful examples and hidden information, unique computational approaches are required. Data mining has, without a doubt, begun to be associated with medical services and clinical data [19].



<Fig. 3> Healthcare data circulation process

3. KNN Models

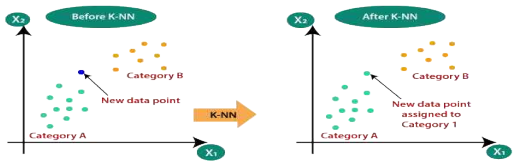
Information is becoming increasingly important in medical care organizations. The fundamental task in clinical science is to determine the cause of any infection and treat patients. In recent years, specialists' manually written notes have been replaced with computerized data, with the goal of lowering treatment costs and increasing treatment effectiveness [18]. Finding and predicting illnesses is critical in the medical care industry, and it is likely the primary motivator

for using data mining for medical care [11], [20]. The use of data mining for medical treatment has aided experts in increasing the well-being administrations they provide [6], [7]. For the order in which a preparation informational index is presented, the K closest neighbor calculation is used. This algorithm uses a search process to select the best fit class for new prepared information from K nearest neighbors of the supplied contribution within a defined distance. [8] When preparing information that is extremely large, it is a more expensive calculation because the entire informational index is examined to find the best closest neighbor. To create a model for this calculation, there is no learning interaction. The data used in the calculation is a model in and of itself

[29], [32]. KNN is one of the most straightforward and simple data mining algorithms. As training examples, it's known as memory-based categorization. When working with continuous attributes, Euclidean distance [31] is used to find the difference among them. When using the Euclidean formula, one of the most common problems is that big values commonly obliterate smaller ones (for example, in patients). The cholesterol measure runs from 100 to 190, whereas the age measure ranges from 40 to 80, therefore the cholesterol measure will have a bigger influence than the age measure while looking for heart disease records. Continuous properties are normalized to have the same effect on the distance between instances to alleviate this problem [22], [27].

Ⅲ. KNN Based Medical Data Analysis

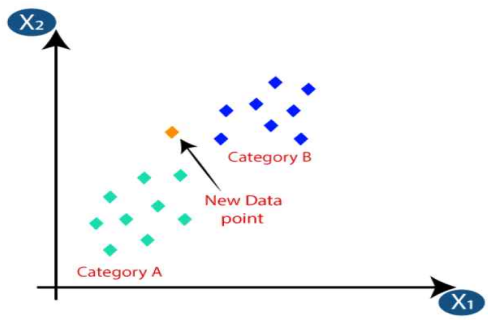
Assuming having two categories (A and B) as indicated in Figure 4, and we have a new data point x1. This data point will fall into which of these groups? A KNN method is use to solve problems like this. This simply determine the category or class of a data set with the help of KNN.



<Fig. 4> Categorization of data set by Euclidean technique

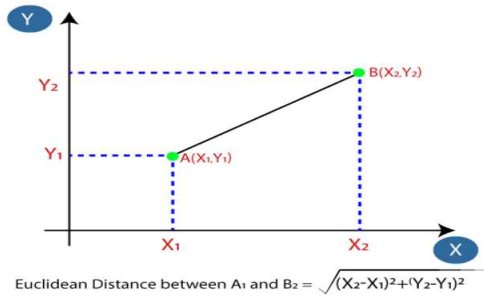
The KNN calculates the Euclidean

distance of "K" number of neighbors, selects the K-nearest neighbors based on the estimated Euclidean distance, then counts the number of data points in each category among these k neighbors [23]. As shown in Figure 5, at the conclusion, assign the new data points to the category with the most neighbors. Let's pretend we have a new data point that needs to be assigned to the best category.



<Fig. 5> Identify the Outlier from both categories

Firstly, we will choose the number of neighbors, so we will choose the K=5. Next, we will calculate the Euclidean distance between the data points (i.e., Euclidean distance is the distance between two points), which we have already studied in geometry.



<Fig. 6> Euclidean distance in different categories

The Euclidean distance was used to find the closest neighbors, yielding three closest neighbors in category A and two closest neighbors in category B. As depicted in the diagram. Because six of the nearest neighbors belong to category A, this new data point must as well, there is no one-size-fits-all method for determining the ideal value for "K," so we'll have to experiment with a variety of options to find the greatest one. The value of K that is most commonly used is 5. K=1 or K=2 is an extremely low value for K that might be noisy

and cause outlier effects in the model. Large values for K are desirable, but they may cause problems. The algorithm calculates the distance between the data to be evaluated and all training data. After then, the distance will be sorted ascending in order K (for example, if K = 10, KNN will sort from 1 to 10). Following that, KNN will pair the relevant data and look for the number of classes from the closest neighbors, after which the class will be designated as the data class to be evaluated. If you enter K = 3, the classification results will be projected as the top three data or data with numbers 1 through 3. The categorization results from the generated user data will then be calculated using the closest class calculated from the three data.

Ⅳ. Conclusion

This article has discussed K-data mining in the clinical setting and the amount of clinical data that can be useful. Within the realm of medical care associations, many data mining applications are split down. This article examines various information mining processes, their benefits and drawbacks, and concludes that there may not be a single information mining strategy that can reliably produce expected results for a wide range of medical services data Without a doubt, the display of tactics varies from one dataset to the next. There is a need to improve and receive well-being information segregated into diverse gathers for compelling use of these approaches in the medical services area [24]. In several commercial fields, data mining is useful for finding examples, determining, and disclosing information, among other things. Data mining has a long range of useability in nearly every business where data is generated, which is why it is regarded as one of the most promising transdisciplinary breakthroughs in information and communication technologies. The K-nearest neighbor approach is a supervised learning methodology for pattern recognition, classification, estimation, and prediction [25]. The memory-based K-nearest neighbor method has a basic premise and does not require a specific model. Other than collecting vectors labeled with the class chosen, no extra training strategy is required for a collection of observations. All intensive

calculations are done on classifications with two observations: finding the closest "K" value in the training set and seeking the most votes in iteration K, as well as class labeling in each classification.

참 고 문 헌

S.D. Gheware, A.S.Kejkar, S.M.Tondare, Data Mining: Task, Tools, Techniques and Applications, International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 10, October 2014. [1]

N. Seth, D. Johnson, G. W. Taylor, O. B. Allen, and H. A. Abdullah, ‘‘Robotic pilot study for analyzing spasticity: Clinical data versus healthy controls,’’ J. Neuroeng. Rehabil. vol. 12, no. 1, 2015, Art. No. 109. [2]

P. T. Noi and M. Kappas, ‘‘Comparison of random forest, K-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery,’’ Sensors, vol. 18, no. 1, p. 18, 2018. [3]

Divya Tomar and Sonali Agarwal, A survey on data mining approaches for healthcare, International Journal of Bio-Science and Bio-Technology Vol.No.5, pp. 241-266, 2013. [4]

G. Beller, J. Nucl. Cardiol. “The rising cost of health care in the United States: is it making the United States globally noncompetitive?” vol. 15, no. 4, pp. 481-482, 2008. [5]

J. Feng, Y. Wei, and Q. Zhu, ‘‘Natural neighborhood-based classification algorithm without parameter k,’’ Big Data Mining Anal., vol. 1, no. 4, pp. 257-265, Dec. 2018. [6]

Gosain, A.; Kumar, A., "Analysis of health care data using different data mining techniques," Intelligent Agent & Multi-Agent Systems, 2009. IAMA 2009, International Conference on, vol. no., pp.1, 6, 22-24 July 2009. [7]

Dr. M.H.Dunham, “Data Mining, Introductory and Advanced Topics”, Prentice Hall, 2002. [8]

Naren Ramakrishnan, David Hanauer, Benjamin J. Keller, Mining Electronic Health Records, IEEE

Computer 43(10): 77-81, 2010. [9]

Soni J, Ansari U, Sharma D, "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887), Volume 17- No.8, March 2011. [10]

Vercellis, C., 2009. Business Intelligence; Data Mining and Optimization for Decision Making. John Wiley & Sons, Ltd., UK. [11]

V. krishnaiah, G. Narsimha, & N. Subhash Chandra, A study on clinical prediction using Data Mining techniques, International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1, 239 248, March 2013. [12]

Divya Tomar and Sonali Agarwal, A survey on data mining approaches for healthcare, International Journal of Bio-Science and Bio-Technology Vol.No.5, pp. 241-266, 2013. [13]

Mohammed Abdul Khalid, Sateesh kumar Pradhan, G.N.Dash, F.A.Mazarbhuiya, A survey of data mining techniques on medical data for finding temporally frequent diseases", International Journal of Advanced Research in Computer and Communication Engineering Vol.2, Issue 12, December 2013. [14]

Obenshain M. K, Application of data mining techniques to healthcare data Infect. Control Hosp. Epidemiol, 25(8):690-695, 2004. [15]

M. H. Dunham, "Data mining introductory and advanced topics", Upper Saddle River, NJ: Pearson Education, Inc., 2003. [16]

A. Shameem Fathima, D. Manimegalai and Nisar Hundewale, A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, ISSN (Online): 1694- 0814, November 2011. [17]

K. Usha Rani, Analysis of Heart Diseases Dataset Using Neural Network Approach, International Journal of Data Mining

&Knowledge Management Process (Ijdkp) Vol.1, No.5, September 2011. [18]

J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, “Neighborhood components analysis,” in Proceedings of the Conference on Neural Information Processing Systems (NIPS ’04), 2004 [19]

Emina Alickovic and Abdulhamit Subasi, Data Mining Techniques for Medical Data Classification, the International Arab Conference on Information Technology (ACIT), 2011. [20]

S. Anto, Dr.S.Chandramathi, Supervised Machine Learning Approaches for Medical Data Set Classification - A Review, IJCST Vol. 2, Issue 4, Oct - Dec 2011. [21]

Goharian & Grossman, Data Mining Classification, Illinois Institute of Technology, 2003. [22]

Kalyani Mali & Samayita Bhattacharya., Soft computing on Medical - Data (SCOM) for a Countrywide Medical System using Data Mining and Cloud Computing Features, Global Journal of Computer Science and Technology Cloud and Distributed, Volume 13 Issue 3 Version 1.0 Year 2013. [23]

N. Chistianini and J. Shawe-Taylor, “An Introduction to Support Vector Machines, and other kernel-based learning methods”, Cambridge University Press, 2000. [24]

N. Cristianini and J. Shawe-Taylor, “An Introduction to Support Vector Machines”, Cambridge University Press, 2000. [25]

G. Ravi Kumar, Dr. G.A.Ramachandra, K.Nagamani, An Efficient Feature Selection System to Integrating SVM with Genetic Algorithm for Large Medical Datasets International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 2, February 2014. [26]

Argyro Kampouraki, Christophoros Nikou, George Manis, "Robustness of Support Vector Machine-based Classification of Heart Rate Signals", Proceedings of the 28th IEEE, EMBS Annual International Conference, New York, USA, Aug30-sep 3,2006, 1995. [27]

Ankita Agarwal, “Secret Key Encryption algorithm using genetic

algorithm", vol.-2, no.-4, ISSN: 2277 128X, IJARCSSE, pp. 57-61, April 2012. [28]

Milovic Boris and Milovic Milan "Prediction and Decision Making in Health Care using Data Mining", International Journal of Public Health Science (IJPHS), 1(2), pp. 69-78, (2012). [29]

K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," Journal of Machine Learning Research, vol. 10, pp. 207-244, 2009 [29]

L. Guoxiang and Q. Zhiheng, "Data Mining Applications in Marketing Strategy," in 2013 Third International Conference on Intelligent System Design and Engineering Applications, (2013), pp. 518-520. [30]

C. S. Ishikiriyama, D. Miro, and C. F. S. Gomes, "Text Mining Business Intelligence: a small sample of what words can say," Procedia Comput. Sci., vol. 55, pp. 261-267, (2015). [31]