Applications of Data Mining in Healthcare and current issues in Ayurvedic Research

Dr Sanjeevani Jadhao, Om Pandey

[Institutional Affiliation(s)]

Author Note

[Include any grant/funding information and a complete correspondence address.]

Abstract

Quantitative and Qualitative analysis of clinical data using Data Mining techniques could unravel hidden medical knowledge by finding correlations, causations, and associations between apparently independent variables. Data Mining techniques have thus demonstrated a wide range of applications in Healthcare systems from the clinical and diagnostic data. In regard to this, we have discussed various disciplines, methods, models, algorithms, and results, and how these techniques would help in performing studies including but not limited to Long-term prospective and retrospective studies, population studies, correlation studies, multicentric, multiracial, phased studies, meta-analysis, pharmacovigilance, etc. on Ayurvedic drugs and methods. We have discussed the applications of Data Mining on healthcare which are being implemented in developed countries. We have also discussed the issues like lack of quality data and record-keeping, and other issues and challenges in conducting ayurvedic studies and how the National Digital Health Blueprint (NDHB) would be a game changer in the current healthcare system.

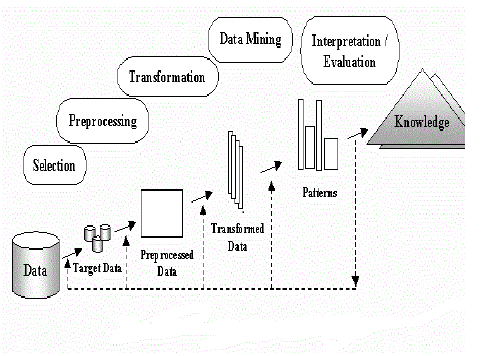
Keywords: Data Mining, Healthcare, Ayurveda, Traditional, NDHB

**Applications of Data Mining in Healthcare and current issues in Ayurvedic Research**

Health and diagnostic data are being collected and accumulated at a vivid pace in the form of Electronic Health Records (EHR) in the developed countries. Data mining has proven effective in areas such as predictive medicine, customer relationship management, detection of fraud and abuse, management of healthcare and measuring the effectiveness of certain treatments. The purpose of data mining, whether it’s being used in healthcare or business, is to identify useful and understandable patterns by analyzing large sets of data. These data patterns help predict industry or information trends, and then determine what to do about them. Different data mining methods have different application in pattern discovery and extraction [1]. The most commonly used data mining techniques are generalization, characterization, classification, clustering, association, evolution, pattern matching, data visualization and meta-rule guided mining [2].

By adopting NDHB and leveraging into AI and ML technologies in a country like India, we will see an evolution in Predictive treatment than a reactive treatment for Ayurvedic medicines. With increased access to a large amount of patient data, healthcare providers would be able to decrease costs by increasing efficiencies, improve patient quality of life, and perhaps most importantly, save the lives of more patients.

# Data Mining: An Overview

Applying computer-based information system (CBIS) to discover knowledge from data is called data mining [3]. Data Mining gives us the ability to extract useful hidden knowledge. With Data Size growing from day to day, this knowledge and information is becoming more important to aid the ayurvedic studies. Data Mining follows 5 major steps i.e., data selection, preprocessing transformation, data mining and Interpretation, and this is how Data is converted into Knowledge.

Data Mining is grounded by the following disciplines-

* Machine Learning
* Artificial Intelligence
* Statistics and Probability

The Interpretation step mainly involves association analysis, Extrapolative modeling, divergence detection etc., to perform these major tasks-

* Classification
* Association rules
* Clustering
* Anomaly detection,

which are achieved with the help of methods like standard support vector data description, density induced support vector data description, Gaussian mixture, vector quantization , statistical, discriminant analysis, decision tree, Markov based, swarm intelligence, k-nearest neighbor, genetic classifiers, artificial neural network, support vector, association rule etc.

## Data Mining Algorithms in Healthcare

Healthcare Analytics covers diagnosis, treatment and, prevention of disease and other physical and mental impairments in humans [4]. Healthcare industry is a data rich industry with massive amount of data in the form Electronic Medical Records (EMR), Claims Data and other benchmarking findings. These data sources are however being under-utilized in the developed countries and rarely utilized in India. Data Mining in Healthcare are mainly used for predicting diseases, making data driven clinical decisions, performing clinical trials and analyzing real world evidence (RWE). As discussed previously, to find valuable information’s in these large volumes of Data using Data Mining, following methods and algorithms are used-

* Anomaly Detection: Anomaly detection is used in discovering the data points which deviates from dataset’s normal behavior [5]. The majorly used anomaly detection methods are, standard support vector data description, density induced support vector data description and Gaussian mixture. The methods are evaluated using the AUC accuracy. The uncertain data points are prone to be available in all datasets, and it could be resolved with anomaly detection.
* Clustering: The clustering is a common descriptive task in which one seeks to identify a finite set of categories or clusters to describe the data [5]. Clustering is approached with vector quantization method to predict the readmissions in intensive medicine. The algorithms used in this method are k-means, k-medoids and x-means. The evaluation is done using Davies-Bouldin Index.
* Discriminant Analysis: Linear discriminant analysis (LDA) is widely used in discriminant analysis to predict the class based on a given set of measurements on new unlabeled observations [5]. The algorithm’s ability to capture statistical dependencies among the predictor variables indicates that this algorithm would be suitable to explore the linear constraint of the studies and the algorithm has good results and it is suitable to be utilized to identify significant accuracy if the relationships of the healthcare data are in linear form.
* Decision Tree: decision tree is useful to improve the prognostic performance, in terms of accuracy because the decision tree recursively separate observations into branches to construct a tree.
* Swarm Intelligence: Swarm intelligence method could be used to design diagnosis model. The algorithm particle swarm optimization (PSO) can efficiently find the optimal or near optimal solutions in large search spaces. The classification process is faster and more accurate if a smaller number of features are used thus improving the overall classification results.
* K-Nearest Neighbor: The k-nearest neighbor is an instance-based classifier method. The parameter units consist of samples that are used in the method and this algorithm then assumes that all instances relate to the points in the n-dimensional space. The algorithm is very effective as the information in the training data is never lost. However, this algorithm would be suitable if the training data set is large as this algorithm is very time consuming when each of the sample in training set is processed while classifying a new data and this process requires a longer classification time. But its classification accuracy is better and it has more significance in medical diagnosis.
* Logistic Regression Logistic regression (LR) is a method that would use the given set of features either continuous, discrete, or a mixture of both types and the binary target, the LR then computes a linear combination of the inputs and passes through the logistic function [6]. This method is commonly used because it is easy to implementation and it provides competitive results. The LR works well for larger datasets
* Bayesian Classifier: The Bayesian classifiers is well known for its computational efficient and ability to handle missing data naturally and efficiently. Having this advantage, it records a good prediction accuracy from the models designed. It is more effective since the averaging approach leads to improved prediction accuracy and allows us to extract more features from the data without being overfitted. This method would be a good approach if the data sets are suffering from missing data.
* Support Vector: The support vector method (SVM) is proven to be advantageous in handling classification tasks with excellent generalization performance. The method seeks to minimize the upper bound of the generalization error based on the structural risk minimization principle. The SVM training is equivalent to solve a linear constrained quadratic programming problem [7]. The method is very commonly used in medical diagnosis. It is also used for comparative study purpose. The SVM method generalization ability is controlled by two different factors, that is the training error and the capacity of the learning machine measured. The training error rate can be controlled by changing the features in the classifiers. SVM shows greater performance since it maps the features to higher dimensional space.

### Applications of Data Mining in Healthcare.

Healthcare industries generate huge amounts of data about patients, hospital resources, clinical trials, disease, diagnosis, electronic patient records, medical devices, Real world evidence etc. Findings and insights from these Data enables support for decision making and cost optimization. Data mining applications in healthcare are as follows-

* **Measuring Treatment effectiveness**: Data mining methods help to evaluate the effectiveness of medical treatments and analyses which course of action proves effective by comparing causes, symptoms and courses of treatments for patients with diverse treatment histories and demographics.
* **Healthcare management**: Data mining methods can identify and track chronic disease, epidemic and pandemic breakouts, high-risk patients etc. and formulate appropriate steps to reduce the impact on the healthcare system and individuals. Data mining could also be applied for early detection and management of pandemics and novel diseases.
* **Public health Policymaking:** By Combining GIS (Geographic Information System) with an integrated database and the application of data mining, the health departments could detect location specific unusual health related events and curtail them before they worsen. This would help in improving the decision making and making necessary policy changes.
* **Evidence-based medications and prevention of hospital errors:** By discovering useful and potentially life-saving insights that otherwise would have remained idle in their databases, safety issues could be flagged and addresses to the regulatory authorities, thus reducing deaths due to hospital errors and human errors.
* **Non-Invasive diagnosis:** Several predictive systems based on clustering methods have proven to show better results than the invasive, costly, and painful diagnosis systems. By analyzing various health attributes, it could be determined if a patient actually needs a biopsy or not. Data driven results have shown higher accuracy than the conventional diagnosis systems.
* **Real World Evidence:** The drugs approved as safe after the Phase 3 clinical trials are regarded as non-harmful but are later found to have adverse effects. US Food and Drug Administration uses data mining to discover knowledge about drug side effects in their database. This algorithm called MGPS or Multi-item Gamma Poisson Shrinker was able to successfully find 67% of ADEs five years before they were detected using traditional ways.
* **Customer relationship management**: Interactions between patients and commercial organizations like bank, insurance, retailers etc. could be initiated and managed by mapping the patients’ expenditure patterns with their financial needs and suggesting patients a suitable health cover accordingly.
* **Fraud and abuse Detection:**  By identifying anomalies and unusual patterns in insurance claims, prescriptions or referrals, the patients or the regulatory authorities could be notified about the fraudulent insurance claims, improper prescriptions or referrals ,thus making it difficult to conduct fraudulent activities in the healthcare systems.
* **Medical Device Industry:** With technological advancements and better understanding of patient needs, One-pass data stream mining algorithms can perform real-time analysis on-board small/mobile devices for a safe and constant way of monitoring of vital signs of patients[8], while considering available resources such as battery charge and available memory thus making medical devices mobile, lightweight, and more effective.
* **Pharmaceutical Industry:** pharmaceutical firms have very recently started managing their inventories and to develop new product and services using Data Mining methods. Understanding of the knowledge hidden in the Pharma data is vital to a firm’s competitive position, understanding prescription patterns and organizational decision-making.
* **Hospital Management:** Temporal and spatial behavior of Hospital activities are globally visualized with the huge amount of hospital data collected. It is further used to optimize Serviced for Hospital Management, Medical staff and Patients.
* **Bioinformatics:** Biological and Genomic data are rich in relational structure hence multi-relational data mining techniques are frequently applied to biological data [9]. Systems biology could be understood better with the health data and hence it’sneeds anddemand has increased in the international sciences.

**Issues and Challenges for Data Mining in Healthcare**

Although Ayurveda literatures contain systematic documentation of clinical experiences, lack of structured data poses a challenge in proving the safety and efficacy of the ayurvedic drugs. The same data could be used to obtain real world evidence and thus improving the credibility of the Ayurveda discipline.

Another issue that comes in picture because of lack of Standardization and Quality Control is that the findings from data mining methods could be incorrect and misleading if the standardization and quality control aren’t looked after.,

The Data Mining methods are good at describing the patterns and trends but not explaining the trends. This failure to be conclusive reduces the credibility of the Data Mining in this segment of healthcare. Moreover, overcoming biases due to majorities and finding minorities that do not confirm to the patterns is another challenge. Furthermore, Ayurveda considers disease occurring due to multifactor and follows multitarget management [ss-Clinical trials in Ayurveda: Issues, challenges and approaches]. Thus, Health Data with limited attributes might not be enough to perform Ayurvedic Studies.

Although the results from data mining methods are credible, the health care practitioners aren’t welcoming enough to accept the results. The Physicians prefer to listen to respected Key Opinion Leaders rather than to the Data Mining results. Thus, convincing the physicians to trust and follow those results would be difficult.

Since individual Health record is sensitive information and multiple parties would be involved in analyzing the data at various steps, ensuring privacy and ethical use of those records would be another big challenge.

##### **Conclusion and Recommendations**

This paper aimed to discuss a brief introduction on Data Mining, Methods used in Healthcare, currently used and Potential application of Data Mining in healthcare system, and Issues and challenges in implementing data mining on ayurvedic studies.

Health care organizations and agencies in countries like India should investigate these applications to find ideas on how to extract knowledge from their own database systems. Healthcare Organization in collaboration with Government of India should start NDHB soon as possible and start leveraging into sophisticated information systems and inculcate these results into the policy making to eliminate issues like infant mortality and disease outbreak (Malaria, Dengue etc.). The same systems could also be used for vaccinations planning, identifying high risk individuals etc. This would also help in establishing the standardization, safety and efficacy benchmarks of the Ayurvedic drugs, rather than just relying on the Empirical evidences and literatures, thus improving the credibility of Ayurveda Discipline. We also need strict laws and clear policies on misuse of the Health records to protect the privacy and security of patient records. Data Mining gives promising results with encouraging accuracies and it would change the landscape of the current Healthcare System

References

[1] R. Agrawal and G. Psaila, “Active data mining,” Current, pp. 3–8, 1995. Last Name, F. M.

**[2] S. H. Liao, P. H. Chu, and P. Y. Hsiao, “Data mining techniques and applications - A decade review from 2000 to 2011,”**

[3] G. E. Vlahos, T. W. Ferratt, and G. Knoepfle, “The use of computer-based information systems by German managers to support decision making,” Inf. Manag., vol. 41, no. 6, pp. 763–779, 2004.

[4] J.-J. Yang, J. Li, J. Mulder, Y. Wang, S. Chen, H. Wu, Q. Wang, and H. Pan, “Emerging information technologies for enhanced healthcare,” Comput. Ind., vol. 69, pp. 3–11, 2015.

[5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, “From data mining to knowledge discovery in databases,” AI Mag., pp.37–54, 1996.

[6] P. J. García-Laencina, P. H. Abreu, M. H. Abreu, and N. Afonoso, “Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values,” Comput. Biol. Med., vol. 59, pp. 125–133, 2015.

[7] B. Zheng, S. W. Yoon, and S. S. Lam, “Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,” Expert Syst. Appl., vol. 41, no. 4 PART 1, pp. 1476–1482, 2014.

[8] Mobile Data Mining for Intelligent Healthcare Support

[9] David Page and Mark Craven, ―Biological Applications of MultiRelationalData Mining‖.