Performance Evaluation of Predictive Data Mining Using Soft Computing Approach

Ms. Meghana A. Deshmukh[1], Prof. S. P. Akarte[2]

[1]*Dept. of Computer Science and Engineering, Prof. Ram Meghe Institute of Technology and Research,Badnera,Amravati,meghnadeshmukh9@gmail.com*
[2]*Dept. of Computer Science and Engineering, Prof. Ram Meghe Institute of Technology and Research, Badnera,Amravati,s_akarte25@rediffmail.com*

Abstract: The paper is about the results on mining of data and finding essential information from large amounts of data. Extracting the knowledge from huge amounts of data is known as Data Mining. Use of algorithms to extract the information and patterns is derived by the KDD process. To finding useful information and patterns in data is Knowledge Discovery in Databases. Research in data mining continues growing in business and in learning organization over coming decades. This report presents a data mining study of medical data with fuzzy. We survey on the theoretical and practical developments of the theory of fuzzy logic and soft computing. Specifically, briefly review the history and main milestones of fuzzy logic (in the wide sense), the more recent development of soft computing, and finalize by presenting a panoramic view of applications.

Keywords: Data Mining, Knowledge Discovery in Databases (KDD). Fuzzy Logic, KNN, Clustering

I. INTRODUCTION

Knowledge Discovery and Data Mining are powerful automated data analysis tools and they are predicted to become the most frequently used analytical tools in the near future [3]. Knowledge discovery and data mining are the landmarks of the information age. Acquiring, storing, and understanding data have posed great challenges and brought a lot of promises. Medical data is often very high dimensional. Depending upon the use, some data dimensions might be more relevant than others. In processing medical data, choosing the optimal subset of features is such important, not only to reduce the processing cost but also to improve the usefulness of the model built from the selected data. This report presents a data mining study of medical data with fuzzy concept that use feature subsets selected by some methods.

Specifically, two fuzzy modeling methods including the fuzzy k-nearest neighbor algorithm, a fuzzy clustering-based modeling are employed databases can contain vast quantities of data describing decisions, performance and operations. In many cases the database contains critical information concerning past business performance which could be used to predict the future. Data mining (also known as Knowledge Discovery) technology helps businesses discover hidden data patterns and provides predictive information which can be applied to benefit the business. The basic approach is to access a database of historical data and to identify relationships which have a bearing on a specific issue, and then extrapolate from these relationships to predict future performance or behaviour.

KDD refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data Prior to the data mining step. Data mining refers to the application of algorithms for extracting patterns from data without the additional steps of the KDD process.

II. TECHNICAL BACKGROUND

Data Mining and Knowledge Discovery in Databases are terms used interchangeably. Other terms often used are data or information harvesting, data archeology, functional dependency analysis, knowledge extraction and data pattern analysis. A high level definition of Data Mining is: the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. Data mining is not a simple process and there is no tool that can do the

job automatically. Data mining can be aided by tools, but it requires both human data mining expertise and human domain expertise. Data mining consists of a number of operations, each of which are supported by a variety of technologies, such as rule induction, neural networks, conceptual clustering [4].
The modeling of imprecise and qualitative knowledge, as well as the transmission and handling of uncertainty at various stages are possible through the use of fuzzy sets. Fuzzy logic is capable of supporting, to a reasonable extent, human type reasoning in natural form. It is the earliest and most widely reported constituent of soft computing. The development of fuzzy logic has led to the emergence of soft computing.

Objectives:

1. Predict the disease with the help of patient's symptoms.
2. The implemented system should provide the preventive measures related to the disease.
3. The implemented application should provide the information about cure time and the cost of the treatment of disease. It will save the time and money of the patient.
4. With the help of this system the patient can get the financial help by providing the information about free medical treatment schemes to them.

III. Data mining and Knowledge discovery in database

A. Knowledge discovery in database:

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.
Knowledge Discovery in Databases brings together current research on the exciting problem of discovering useful and interesting knowledge in databases [6].

1) **KDD process**:

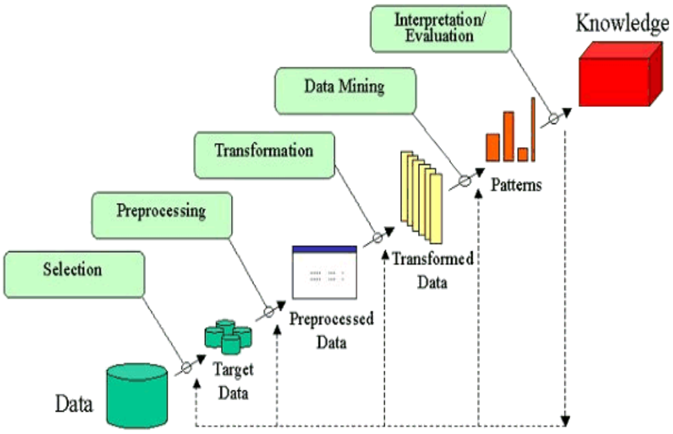


Fig. 1. KDD process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. Brachman and Anand (1996) give a practical view of the KDD process, emphasizing the interactive nature of the process.

Steps are:

- Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

- Data integration: at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

- Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results [6].

It is common to combine some of these steps together. For instance, *data cleaning* and *data integration* can be performed together as a pre-processing phase to generate a data warehouse. *Data selection* and *data transformation* can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

B. **Data mining**

Data mining derives its name from the similarities between searching for valuable information in a large database and mining rocks for a vein of valuable ore. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. Nevertheless, data mining became the accepted customary term, and very rapidly a trend that even overshadowed more general terms such as knowledge discovery in databases (KDD) that describe a more complete process. Other similar terms referring to data mining are: data dredging, knowledge extraction and pattern discovery.

1) **Working of data mining**:

Data mining is an iterative process that typically involves the following phases:

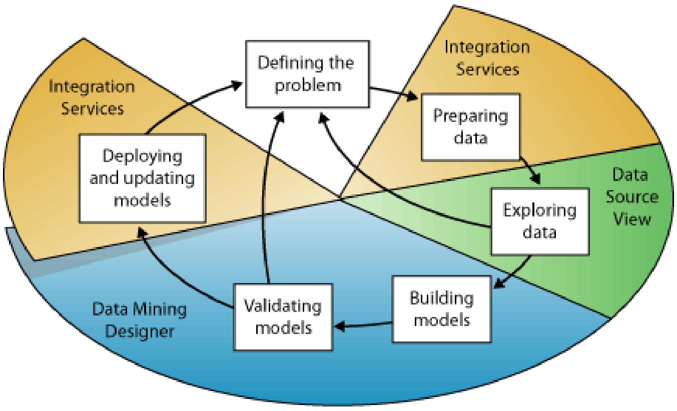


Fig. 2. Data mining process

Problem definition: A data mining project starts with the understanding of the business problem. Data mining experts, business experts, and domain experts work closely together to define the project objectives and the requirements from a business perspective. The project objective is then translated into a data mining problem definition. In the problem definition phase, data mining tools are not yet required [1].

Data exploration: Domain experts understand the meaning of the metadata. They collect, describe, and explore the data. They also identify quality problems of the data. A frequent exchange with the data mining experts and the business experts from the problem definition phase is vital. In the data exploration phase, traditional data analysis tools, for example, statistics, are used to explore the data [1].

Data preparation: Domain experts build the data model for the modeling process. They collect, cleanse, and format the data because some of the mining functions accept data only in a certain format. They also create new derived attributes, for example, an average value. In the data preparation phase, data is tweaked multiple times in no prescribed order. Preparing the data for the modeling tool by selecting tables, records, and attributes, are typical tasks in this phase. The meaning of the data is not changed [1].

Modeling: Data mining experts select and apply various mining functions because you can use different mining functions for the same type of data mining problem. Some of the mining functions require specific data types. The data mining experts must assess each model. In the modeling phase, a frequent exchange with the domain experts from the data preparation phase is required. The modeling phase and the evaluation phase are coupled. They can be repeated several times to change parameters until optimal values are achieved. When the final modeling phase is completed, a model of high quality has been built [1].

Evaluation: Data mining experts evaluate the model. If the model does not satisfy their expectations, they go back to the modeling phase and rebuild the model by changing its parameters until optimal values are achieved. When they are finally satisfied with the model, they can extract business explanations and evaluate the following questions: 1. Does the model achieve the business objective? 2. Have all business issues been considered? At the end of the evaluation phase, the data mining experts decide how to use the data mining results. Deployment Data mining experts use the mining results by exporting the results into database tables or into other applications, for example, spreadsheets [1].

2) Issues in Data Mining

Security and social issues: Security is an important issue with any data collection that is shared and/or is intended to be used for strategic decision-making. In addition, when data is collected for customer profiling, user behaviour understanding, correlating personal data with other information, etc., large amounts of sensitive and private information about individuals or companies is gathered and stored. This becomes controversial given the confidential nature of some of this data and the potential illegal access to the information. Moreover, data mining could disclose new implicit knowledge about individuals or groups that could be against privacy policies, especially if there is potential dissemination of discovered information. Another issue that arises from this concern is the appropriate use of data mining. Due to the value of data, databases of all sorts of content are regularly sold, and because of the competitive advantage that can be attained from implicit knowledge discovered, some important information could be withheld, while other information could be widely distributed and used without control [6].

User interface issues: The knowledge discovered by data mining tools is useful as long as it is interesting, and above all understandable by the user. Good data visualization eases the interpretation of data mining results, as well as helps users better understand their needs. Many data exploratory analysis tasks are significantly facilitated by the ability to see data in an appropriate visual presentation. There are many visualization ideas and proposals for effective data graphical presentation. However, there is still much research to accomplish in order to obtain good visualization tools for large datasets that could be used to display and manipulate mined knowledge. The major issues related to user interfaces and visualization are "screen real-estate", information rendering, and interaction. Interactivity with the data and data mining results is crucial since it provides means for the user to focus and refine the mining tasks, as well as to picture the discovered knowledge from different angles and at different conceptual levels [6].

Mining methodology issues: These issues pertain to the data mining approaches applied and their limitations. Topics such as versatility of the mining approaches, the diversity of data available, the dimensionality of the domain, the broad analysis needs, the assessment of the knowledge discovered, the exploitation of background knowledge and metadata, the control and handling of noise in data, etc. are all examples that can dictate mining methodology choices. For instance, it is often desirable to have different data mining methods available since different approaches may perform differently depending upon the data at hand. Moreover, different approaches may suit and solve user's needs differently.

Most algorithms assume the data to be noise-free. This is of course a strong assumption. Most datasets contain exceptions, invalid or incomplete information, etc., which may complicate, if not obscure, the analysis process and in many cases compromise the accuracy of the results. As a consequence, data preprocessing (data cleaning and transformation) becomes vital. It is often seen as lost time, but data cleaning, as time-consuming and frustrating as it may be, is one of the most important phases in the knowledge discovery process. Data mining techniques should be able to handle noise in data or incomplete information.

More than the size of data, the size of the search space is even more decisive for data mining techniques. The size of the search space is often depending upon the number of dimensions in the domain space. The search space usually grows exponentially when the number of dimensions increases. This is known as the *curse of dimensionality*. This "curse" affects so badly the performance of some data mining approaches that it is becoming one of the most urgent issues to solve [6].

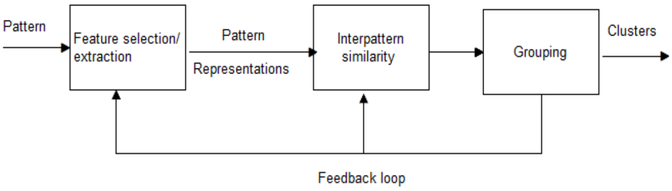
Performance issues: Many artificial intelligence and statistical methods exist for data analysis and interpretation. However, these methods were often not designed for the very large data sets data mining is dealing with today. Terabyte sizes are common. This raises the issues of scalability and efficiency of the data mining methods when processing considerably large data. Algorithms with exponential and even medium-order polynomial complexity cannot be of practical use for data mining. Linear algorithms are usually the norm. In same theme, sampling can be used for mining instead of the whole dataset. However, concerns such as completeness and choice of samples may arise. Other topics in the issue of performance are *incremental updating*, and parallel programming. There is no doubt that parallelism can help solve the size problem if the dataset can be subdivided and the results can be merged later. Incremental updating is important for merging results from parallel mining, or updating data mining results when new data becomes available without having to reanalyze the complete dataset [6].

Data source issues: Verious issues related to the data sources, some are practical such as the diversity of data types, while others are philosophical like the data glut problem. We certainly have an excess of data since we already have more data than we can handle and we are still collecting data at an even higher rate. If the spread of database management systems has helped increase the gathering of information, the advent of data mining is certainly encouraging more data harvesting. The current practice is to collect as much data as possible now and process it, or try to process it, later. The concern is whether we are collecting the right data at the appropriate amount, whether we know what we want to do with it, and whether we distinguish between what data is important and what data is insignificant. Regarding the practical issues related to data sources, there is the subject of heterogeneous databases and the focus on diverse complex data types. We are storing different types of data in a variety of repositories. It is difficult to expect a data mining system to effectively and efficiently achieve good mining results on all kinds of data and sources. Different kinds of data and sources may require distinct algorithms and methodologies. Currently, there is a focus on relational databases and data warehouses, but other approaches need to be pioneered for other specific complex data types. A versatile data mining tool, for all sorts of data, may not be realistic. Moreover, the proliferation of heterogeneous data sources, at structural and semantic levels, poses important challenges not only to the database community but also to the data mining community [6].

- **Clustering**

Before each application of collaborative filtering, clustering is applied to the training set to discover connected components of patients. This served to remove the inuence of patients who have little or no similarity with the testing patient for whom predictions are being made. This is determined by the number of diseases which the patients have in common. In the most basic case, patients are removed only if they have no diseases in common with the active patient [9].

Based on the above, a unified theoretical framework for data mining is presented by formulating a unified data mining theory (UDMT) where the data mining processes; clustering, classification and visualization are unified by means of composition of functions. The proposed unified theoretical framework is based on the following assumptions which are also called the steps for knowledge extraction from a dataset:
Step 1: Create (appropriate) partitions of the dataset.
Step 2: Create the clusters of each partition. This step is also called the clustering. The clustering is a technique of dividing a dataset into different groups. The goal of clustering is to find groups that are very different from each other and whose members are very similar to each other.
Step 3: Construct KNN on each clusters. This step is also referred as the classification. The classification is a technique of placing an object into group based on common properties among the objects.
Step 4: Plot the 2D or 3D graphs of each classifier. This step is also known as visualization. The visualization is a process of presenting data in a special and easy to understandable form. It is also a relationship within the data which is not evident from the raw data.



- **Fuzzy k-nearest neighbor**

KNN algorithm is a Lazy Learning Algorithm. Defer the decision to generalize beyond the training

examples till a new query is encountered. Whenever we have a new point to classify, we find its K nearest neighbors from the training data. The fuzzy k-nearest neighbor (KNN) algorithm was proposed by Keller. As a generalization of KNN algorithm to allow the assignment of fractional membership, instead of zero or one like KNN, to each class. fuzzy KNN first uses a training dataset to obtain membership values of how training vector instances belong to each class. After training, fuzzy KNN can determine the class of a test instance based on the distances from its nearest neighbors and each neighbor's membership in each class. Generally, the test instance is assigned the class that is represented by the majority of its k nearest neighbors in the training dataset [4].

IV. Proposed System Planning

- Disease Pattern Matching application using fuzzy set logic is used.

- With the help of this dissertation, the information regarding the Disease can be gained.

In this dissertation, the framework proposed is applied to additional applications for a variety of sequential patterns discovering based on fuzzy concept. The framework also can extend and examine to discover association rules with fuzzy concept in the future.
By providing a research result to demonstrate that this framework can be a generalization of this realm problem. Discovering sequential patterns can reveal what items are frequently bought together and in what order they appear . Although that sort of patterns can provide lots of interesting information, they cannot tell us the time gaps between successive items.

V. Results

Disease name	Worst Case	Best Case
food poisoning	11 ms	01 ms
Alopecia	09 ms	03 ms
Acute cough	11 ms	04 ms

Dengue	02 ms	01 ms
Ache, ear	12 ms	03 ms
difficulty sleeping	13 ms	03 ms
Glossitis	06 ms	01 ms

Table 1 : Best case & Worst case result

Disease Name	Frequency
food poisoning	2
Alopecia	5
Acute cough	4
Dengue	3
Ache, ear	1
difficulty sleeping	5
Glossitis	2

Table 2 : Frequency of Disease

VI. CONCLUSION & FUTURE SCOPE

To aid clinicians in the diagnosis of disease, recent research has looked into the development of computer aided diagnostic tools. Various data mining techniques have been widely used for breast cancer diagnosis heart disease or any other disease. Current research in data mining mainly focuses on the discovery algorithm and visualization techniques. There is a growing awareness that, in practice, it is easy to discover a huge number of patterns in a database where most of these patterns are actually obvious, redundant, and useless or uninteresting to the user. To prevent the user from being overwhelmed by a large number of uninteresting patterns, techniques are needed to identify only the useful/interesting patterns and present them to the user. Soft computing methodologies, involving fuzzy sets, neural networks, genetic algorithms, rough sets, and their hybridizations, have recently been used to solve data mining problems. They strive to provide approximate solutions at low cost, thereby speeding up the process. A categorization has been provided based on the different soft computing tools and their hybridizations used, the mining function implemented, and the preference criterion selected by the model. Fuzzy sets, which constitute the oldest component of soft computing, are suitable for handling the issues related to understandability of patterns, incomplete/noisy data, mixed media information and human interaction, and can provide approximate solutions faster. They have been mainly used in clustering, discovering association rules and functional dependencies, summarization, time series analysis, web applications and image retrieval. This system predicts the any disease with respect to the symptoms. The system also provides the another benefits. this system have the potential to show the disease related suggested reports and the patients can get the financial help with this system, the system provides the information about government free treatment schemes. The system has fulfilled all the objectives that mentioned above.

In future the predictor can be used to design a web based application. By using the another soft computing techniques it can be possible to make this project very large and huge. Diagnosis can be done in any specific area or by adding some more information in this project. fuzzy concept and Clustering is also found to be a popular technique in medical prediction Particular it has been successfully utilized for diagnosis. In future we intend to design and implement such system for web based applications. The text mining can be used to mine huge amount of unstructured data available in healthcare industry database.

REFERENCES

[1] “Data Mining, Applications and Knowledge Discovery” International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970) Volume-2 Number-4 Issue-6 December-2012.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth "From Data Mining to Knowledge Discovery in Databases".

[3] "On the Impact of Knowledge Discovery and Data Mining" crpit.com/confpapers/CRPITV1Wahlstrom.pdf.

[4] http://www.aiai.ed.ac.uk/links/dm.html#intro.

[5] seclab.cs.ucdavis.edu/projects/misuse/meetings/*KDD.html*.KDD Overview Notes.

[6] http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/notes/Chapter1/."Introduction to Data Mining".

[7] Agrawal, R., and Psaila, G. 1995. Active Data Min- ing. In Proceedings of the First International Con- ference on Knowledge Discovery and Data Mining (KDD-95), 3–8. Menlo Park, Calif.: American Asso- ciation for Artificial Intelligence.

[8] "Introduction to Data Mining and Knowledge Discovery" by Two Crows Corporation. www.twocrows.com/intro-dm.pdf.

[9] Tipawan Silwattananusarn and Assoc.Prof. Dr. Kulthida Tuamsuk, "Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012", International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol.2, No.5, Thailand, September 2012.

[10] "The KDD Process for Extracting Useful"**.** Knowledge from Volumes of Data. shawndra.pbworks.com/.../.."