CHANGING PARADIGM OF DATA MINING

Vishal Sharma

Assisstant Professor, Department Of Computer Science and IT,
DAV College, Jalandhar
vishalorsharma@gmail.com

Abstract

Data mining is an ever evolving technique with immense potential to help modern day business houses, focus on the most significant information in their data warehouses, by the extraction of hidden predictive information from large databases. With the exponential growth of data in terms of volume, velocity and veracity, the development of data mining tools has gained immense significance. The automated, prospective analysis offered by data mining move beyond the analysis of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time-consuming to resolve. Data mining techniques can be implemented rapidly on existing software and hardware platforms to enhance the value of existing information resources. The classical techniques of statistics and regression can be integrated with modern soft computing technologies to predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions so as to stay in contention.

Keywords: *Data Mining, Data warehouses, Knowledge- driven decision, Predictive information, Soft Computing.*

1. Introduction: Data Mining

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time [1]. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining has enormous application in the business community because it is supported by three technologies that have grown sufficiently mature over time:

- Massive data collection and storage technology
- Powerful multiprocessor computers
- Data mining algorithms and techniques

With the growth of commercial databases at unprecedented rates, some industries, particularly retail (Amazon, Alibaba, Flipkart, Walmart etc.) the numbers can be much beyond anticipation. Data mining algorithms embody techniques that may have existed for past many years, but have only recently been implemented as mature, reliable, understandable tools that have consistently outperformed older statistical methods. A new technological leap is needed to structure and prioritize information for specific end-user problems [2].

In the transformation of business data to business information, each new step was built upon the previous one. The four steps [3] listed in ***Table 1*** are revolutionary because they have allowed new business questions and strategies to be answered accurately and in quick time.

Today, the maturity of these techniques, coupled with high-performance relational database engines and broad data integration efforts, make them relevant for current data warehouse environments.

Evolutionary Step	**Business Question**	**Enabling Technologies**	**Product Providers/Tools**	**Characteristics**
Data Collection (1960s)	"What was my companie's total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective, static data delivery
Data Access (1980s)	"What were unit sales in New England last March?"	Relational databases Structured Query Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Evolving Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessing, massive databases, machine learning, plus- R, Python, ANN, Fuzzy logic	Pilot, Lockheed, IBM, SGI, Nascent industry, WEKA, Rapid Miner, Fuzzytech, R Studio, Orange, Matlab	Prospective, proactive information delivery

Table 1 Steps in the Evolution of Data Mining[1,2,4,5,6]

2. Data Mining Techniques

Data mining is the process of extracting patterns from data. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage [7]. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery and retails in recent times. This section provides a description of some of the most common data mining algorithms and techniques being used in today's world.

2.1 Classical Techniques: Statistics, Neighborhoods and Clustering

A number of proprietary techniques from particular vendors as well as many others in general are available today, however the industry is converging to those techniques that are consistent, understandable and comprehendible. Some of them are:

2.1.1 Statistics

Statistical techniques or "statistics" do not translate to data mining; however they had been used long before the term data mining was associated to business applications. The statistical techniques are driven by the data and are used to discover patterns and build predictive models. In statistics, prediction is usually synonymous with regression of some form. A variety of regression techniques (linear, logistic, polynomial, stepwise etc) exist, but the basic idea is to create a model that maps values from predictors to minimize error in making a prediction. These techniques are mostly driven by three metrics (number of independent variables, type of dependent variables and shape of regression line) as shown in fig 1(a). Regression analysis is an important tool for modeling and analyzing data. For linear regression the line takes a given value for a predictor and map it into a given value for a prediction (Prediction = a + b * Predictor), fig 1(b). The aim is to fit a curve / line to the data points, fig 1(c) in such a manner that the differences between the distances of data points from the curve or line is minimized as illustrated fig 1(d).

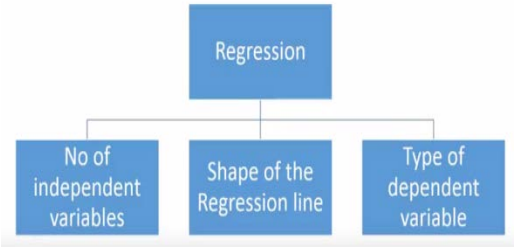

Fig 1(a): Regression Metrics[29]

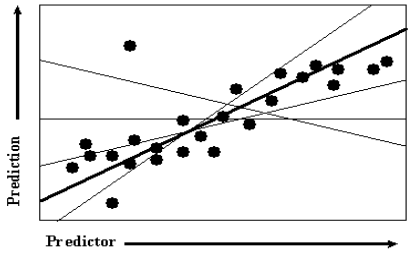

Fig 1(b): Linear Model of Regression

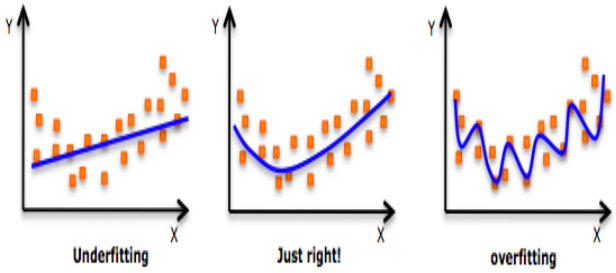

Fig 1(c): Model Selection[29]

Fig 1(d): Ideal Model of Regression[29]

The choice of technique is driven by the types of independent and dependent variables, dimensionality of data, probability of event, multicollinearity and other essential characteristics of the data .Objective of the project, statistical significance and cross-validation(division of data set into two group (train and validate) can serve as the best ways to evaluate models used for prediction.

2.1.2 Neighborhoods and Clustering:

2.1.2.1 K-Nearest Neighbors

Nearest Neighbor prediction technique is among the oldest techniques used in data mining. Most people have an intuition that they understand what clustering is - like records are grouped or clustered together [9]. Nearest neighbor has a similarity to clustering in essence, as in order to predict what a prediction value is in one record, look for records with similar predictor values in the historical database and use the prediction value from the record that it "nearest" to the unclassified record, Fig 2(a). One of the improvements to the basic nearest neighbor algorithm is to take a vote from the "K" nearest neighbors rather than just relying on the sole nearest neighbor to the unclassified record. KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry.

Let's take a simple case to understand this algorithm. Following is a spread of red circles (RC) and green squares (GS). To determine the class of the blue star (BS), BS can either be RC or GS and nothing else. The "K" in KNN algorithm is the number of nearest neighbors whose vote counts for. Let's say K = 3. Hence, we will now make a circle with BS as center just as big as to enclose only three data points on the plane. The three closest points to BS is all RC which suggests that BS should belong to the class RC.

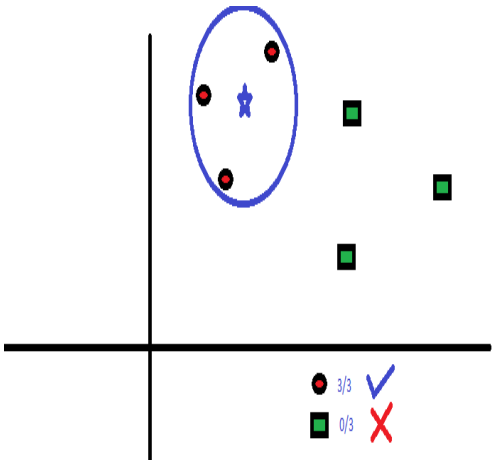

Fig 2(a): Classification of unknown[29]

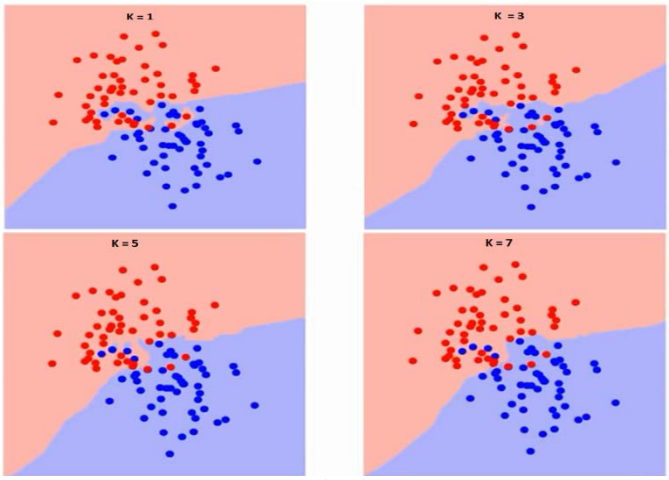

Fig 2(b): Optimal value of K for class boundaries[29]

With a given K-value boundaries of each class can be established which will segregate RC from GS. The same way, the effect of value "K" on the class boundaries can be studied in Fig 2(b) the different boundaries separating the two classes with different values of K can be seen. The boundary becomes smoother with increasing value of K. With K increasing to infinity it finally becomes all blue or all red depending on the total majority. The training error rate and the validation error rate are two parameters needed to access on different K-values. To get the optimal value of K, segregation of the training and validation from the initial dataset is essential. Usually this is accomplished by simply taking the majority of predictions from the K nearest neighbors if the prediction column is a binary or categorical or taking the average value of the prediction column from the K nearest neighbors.

2.1.2.2 Clustering

Cluster analysis or Clustering is the grouping of a particular set of objects based on their characteristics and aggregating them according to their similarities. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. In the context of data mining, hard partitioning, partitions the data to implement a specific join algorithm, suitable for the desired information analysis allowing an object to strictly belong to a cluster or not to be part of it. On the other hand, soft partitioning states that every object belongs to a cluster with a determined degree. Broadly clustering can be of 2 types, partitional clustering: dividing data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset or hierarchical clustering: nested clusters organized as a hierarchical tree. A variety of clusters can be established on the basis of clustering algorithms/techniques and the choice depends on characteristics of the data set and what is desired. (Table 2) [10,11].

Type	Criteria/Characteristics
Well-Separated Clusters	Any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.
Centroid-based	Object in a cluster is closer (more similar) to the "center" of a cluster, than to the center of any other cluster.
Contiguous Cluster (Nearest neighbor or Transitive)	A point in a cluster is closer (or more similar) to one or more other points in the cluster than to any point not in the cluster.
Density-based	A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density used when the clusters are irregular or intertwined, and when noise and outliers are present.
Conceptual Clusters	Clusters share some common property or represent a particular concept. .
Objective Function	Clusters minimize or maximize an objective function(local or global)

Table 2: Types of Clusters/ Clustering [10, 11]

2.2 Next Generation Techniques: Decision Trees, Neural Networks, Fuzzy Logic

These data mining techniques represent the most often used techniques that have been developed over the last two decades of research. These techniques can be used for either discovering new

information within large databases or for building predictive models [12]. Though the older decision tree techniques such as CHAID are currently highly used the new techniques such as CART, Neural Networks and Fuzzy logic and Hybrid approaches are gaining wider acceptance.

2.2.1 Decision Trees

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node. A decision tree is a predictive model for a concept, for instance ***buy_computer*** that indicates whether a customer at a company is likely to buy a computer or not (Figure 3) [13].

A decision tree is used in data mining to simplify complex strategic challenges and evaluate the cost effectiveness of research and business decisions. The benefits of having a decision tree are minimal requirement of any domain knowledge, ease of comprehension, simplicity and quick learning and classification steps. ***Pre-tree pruning,*** halting tree construction early or ***Post-tree pruning***, removing a sub-tree from a fully grown tree are performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex. Cost complexity is measured by the number of leaves in the tree, and error rate of the tree.
Some significant decision tree types/ techniques **ID3** (Iterative Dichotomiser) and **C4.5**, **CHAID** (Chi-Square Automatic Interaction Detector), **CART** (Classification and Regression Trees) - Growing a forest and picking the best tree and their approaches in building a decision tree are summarized in Table 4.

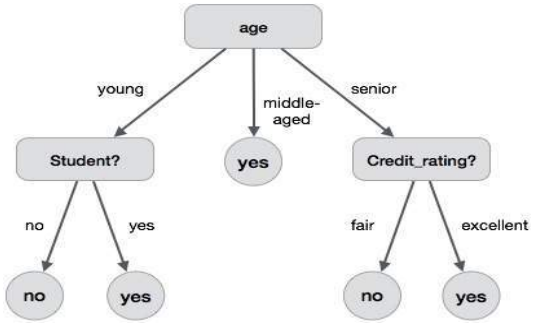

Figure 3: A decision tree as a predictive model[13]

Type	Approach and characteristics
ID3 and C 4.5	Greedy, top-down recursive divide and conquer, no backtracking
CHAID	Uses categorical predictors, coercion by binning[14] Splitting through contingency tables[15]
CART	Data exploration and prediction, generate optimal split even with missing data[15]

Table 4: Decision Trees and Characteristics

2.2.2 Artificial Neural Networks (ANN)

Neural Networks are analytic techniques based on the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain, making them capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after

executing a process of so-called *learning* from existing data which makes it an important data mining technique. The first step is to design a specific network architecture that includes a specific number of "layers" each consisting of a certain number of "neurons", Figure 4. The size and structure of the network needs to match the nature (e.g., the formal complexity) of the investigated phenomenon. Because the latter is obviously not known very well at this early stage, this task is not easy and often involves multiple "trials and errors", however, neural network software exist that applies artificial intelligence techniques to aid in that tedious task and finds "the best" network architecture [16,17]

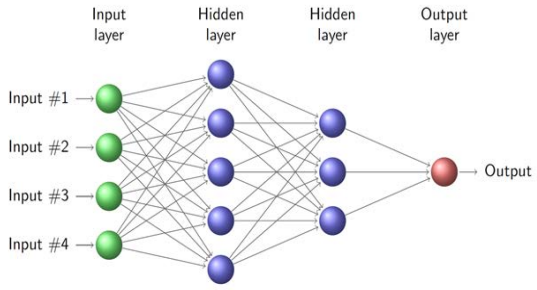

Figure 4: Multilayered Neural Network

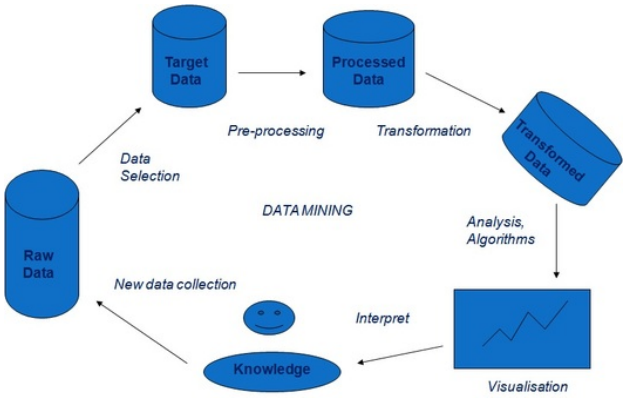

Figure 5: Fuzzy Logic in Data Mining[28]

The greatest breakthroughs in neural network in recent years are in their applications to real world problems like customer response prediction, fraud detection etc. They model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications [18]. ANNs have become a powerful tool in tasks like classification, clustering, feature mining, pattern recognition, and decision problem or prediction applications. ANN is an adaptive, non linear system that learns to perform a function from data (training phase) where system parameters change during operations. After the training is complete the parameter are fixed. The non linear characteristics of ANN model provide good flexibility to achieve input output map for huge volumes of data and poorly understandable problem with great accuracy by selecting network topology, performance parameters, learning rule and stopping criteria.

The neural network model can be broadly divided into three types: feed- forward, feed-back, and self-organization. ANN has the capability of rapidly fitting the non-linear data; therefore it can solve many problems which are difficult for other methods to solve. Neural networks as data mining tools aid in ***data preparation*** involving cleaning selecting, preprocessing and expression of data, ***rule extraction*** by using black-box, fuzzy and recursive networks and finally ***rule assessment*** on the basis of testing the optimality and accuracy of rules extracted, amount of knowledge not extracted

and inconsistency between extracted rules and trained neural network. Out of a sizeable number of neural networks used for data mining, two types are most significant: self organizing neural network and fuzzy neural network [19].

2.2.3 Fuzzy Logic

Fuzzy logic deploys the concept of fuzzy sets and gradual membership to categorize a concept in an abstract way by introducing vagueness. On the contrary, data mining methods can extract patterns automatically from a large amount of data. The integration of fuzzy logic with data mining methods thus creates more abstract patterns at a higher level than at the data level [19]. The business understanding and data understanding phases are usually strongly human centered and only little automation can be achieved here. These data preparation phases serve mainly to define the goals of the knowledge discovery project, to estimate its potential benefit, and to identify and collect the necessary data, background domain knowledge and meta-knowledge. Fuzzy set methods (e.g fuzzy clustering the data to detect outliers) become an obvious choice to formulate the background domain knowledge in vague terms, that can still be used in a subsequent modeling phase. The modeling phase, in which models are constructed from the data in order, for instance, to predict future developments or to build classifiers, can of course, benefit most from fuzzy data analysis approaches(analyze fuzzy data, fuzzy clustering and neuro-fuzzy systems). In the evaluation phase, in which the results are tested and their quality assessed, the usefulness of fuzzy modeling methods becomes most obvious to yield interpretable systems, they can easily be checked for plausibility against the intuition and expectations of human experts (Figure 5). In addition, the results can provide new insights into the domain under consideration, in contrast to, e.g., pure neural networks, which are black boxes [20]. Fuzzy systems adhere well with data mining for the induction of fuzzy rules in order to interpret the underlying data linguistically. To describe a fuzzy system completely we need to determine a rule base (structure) and fuzzy partitions (parameters) for all variables. The data driven induction of fuzzy systems by simple heuristics based on local computations is usually called neuro-fuzzy [21].

Fuzzy techniques for data mining bear an immense significance in real-world applications in terms of: the notion of ***similarity*** and the ***fuzzy machine learning*** techniques. S***imilarity,*** or more generally comparison measures are used at all levels of the data mining and information retrieval tasks. ***Fuzzy Machine Learning*** uses the previous similarity measures and is used as an important way to extract knowledge from sets of cases, especially in large scale databases. The three most widely used methods are ***fuzzy decision trees, fuzzy prototypes and fuzzy clustering***. The first two belong to the supervised learning framework, whereas fuzzy clustering belongs to the unsupervised learning framework, i.e. no a priori decomposition of the data set into categories is available.

a. ***Fuzzy Decision Trees***: Fuzzy decision trees (FDT) have a great affinity for data mining and information retrieval because they enable the user to take into account imprecise descriptions of the

cases, or heterogeneous values (symbolic, numerical, or fuzzy). Moreover, they are appreciated for their interpretability, because they provide a linguistic description of the relations between descriptions of the cases and decision to make or class to assign. The rules obtained through FDT make it easier for the user to interact with the system or the expert to understand, confirm or amend his own knowledge. Another quality of FDT is their robustness, since a small variation of descriptions does not drastically change the decision or the class associated with a case [22].

b. ***Fuzzy Prototype Construction***: Fuzzy prototype constitute another approach to the characterization of data categories: they provide descriptions or interpretable summarizations of data sets, so as to help a user to better apprehend their contents: a prototype is an element chosen to represent a group of data, to summarize it and underline its most characteristic features [23,24,25]. It can be defined from a statistic point of view, for instance as the data mean or the median; more complex representatives can also be used, as the Most Typical Value [26] for instance. The prototype notion from a cognitive science point of view, emphasize the distinctive features as opposed to other categories, underlining the specificity of the group. Furthermore, prototypes were related to the typicality notion: resemblance to other members of the group (internal resemblance), and on its dissimilarity to members of other groups (external dissimilarity) [27].

c. ***Fuzzy Clustering***: Fuzzy clustering is the process of dividing data elements into classes or clusters so that items in the same class are as similar as possible, and items in di fferent class dissimilar as possible. Depending on the nature of the data and the purpose for which clustering is being used; di fferent meas place items into classes. In fuzzy clustering, data elements can belong to more than one cluster, and associated with each element is a set of membership level which indicates the strength of the association between that data element and a particular cluster. Clustering takes place in the product space of systems inputs and outputs and each cluster corresponds to a fuzzy IF-THEN rule. Starting with a number of clusters and subsequently removing less important ones as the clustering progresses, it is sought to obtain a suitable partition of the data in an automated manner. Some of the most widely used fuzzy clustering algorithms are Fuzzy C-means, Fuzzy K-means, and adaptive fuzzy clustering finding applications in bioinformatics, medicine, image processing and marketing.

3. Conclusion

With the regard to moving time, competitiveness, evolving business scenario, diversity, target customers, and strategic improvisation, data mining has evolved over time. Comprehensive data warehouses that integrate operational data with customer, supplier, and market information have resulted in an explosion of information. Competition requires timely and sophisticated analysis on an integrated view of the data. However, there is an ever increasing demand to bridge the growing gap between more powerful storage and retrieval systems and the users' ability to effectively analyze and act on the information they contain. The classical relational and OLAP technologies have tremendous capabilities for navigating massive data warehouses, but brute force navigation of data is

not enough and a technological leap is constantly needed to structure and prioritize information for specific end-user problems [2].

There are definite differences in the types of problems that are most conducive to each technique but the reality of real world data and the dynamic way in which markets, customers and hence the data that represents them is formed means that the data is constantly changing [17]. It no longer makes sense to build the "perfect" model on the historical data since whatever was known in the past cannot adequately predict the future because the future is so unlike what has gone before. The study indicates that varied approaches of data mining are suitable for diverse problem domains, therefore the classical techniques of data-mining need to work in conjunction with modern soft computing techniques in order to meet the dynamic needs of industry.

4. References

[1] Gartner Group Advanced Technologies and Applications Research Note, 1995.

[2] META Group Application Development Strategies: “Data Mining for Data Warehouses: Uncovering Hidden Patterns.”, 1995

[3] http://www.thearling.com/text/dmwhite/dmwhite.htm

[4] Prasad MC, Florence L, Arya A., “A Study on Software Metrics based Software Defect Prediction using Data Mining and Machine Learning Techniques”, International Journal of Database Theory and Application,8(3), 179-190, 2015

[5] Zadeh, Lotfi A., “Soft computing and fuzzy logic”, IEEE software, 11(6), 48, 1994.

[6] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH., “The WEKA data mining software: an update”, ACM SIGKDD explorations newsletter, 11(1), 10-18, 2009.

[7] Gartner Group High Performance Computing Research Note, 1995.

[8] Forrester Research 2010 Predictive Analytics and Data Mining Solutions report.

[9] Annual Rexer Analytics Data Miner Surveys.

[10] www.users.cs.umn.edu/~kumar001/dmbook/dmslides/chap8_basic_cluster_analysis.pdf

[11] http://bigdata-madesimple.com/what-is-clustering-in-data-mining

[12] Robert Nisbet's 2006 Three Part Series of articles Data Mining Tools: “Which One is best for CRM?".

[13] https://www.tutorialspoint.com/data_mining/dm_dti.htm

[14] Haughton et al.'s 2003 “Review of Data Mining Software Packages in The American Statistician”.

[15] Bradley, I., Introduction to Neural Networks, Multinet Systems Pvt Ltd 1997.

[16] Haykin, S., *Neural Networks*, Prentice Hall International Inc., 1999.

[17] Khajanchi, Amit, Artificial Neural Networks: The next intelligence.

[18] R. Andrews, J. Diederich, A. B. Tickle, ”A survey and critique of techniques for extracting rules from trained artificial neural networks”, Knowledge-Based Systems,Vol.- 8, No.-6, pp. 378-389,1995).

[19] Kanhaiya Lal, N.C.Mahanti, "Role of soft computing as a tool in data mining". International Journal of Computer Science and Information Technologies, Vol. 2 (1), 2011, 526-537)
[20] Rudolf Kruse, Detlef Nauck , Christian Borgelt, "Data Mining with Fuzzy Methods": Status and Perspectives", In Proceedings of the EUFIT, 99.
[21] D. Nauck, F. Klawonn, and R. Kruse. Foundations of Neuro-Fuzzy Systems. J. Wiley & Sons, Chichester, United Kingdom 1997)
[22] M. Rekha, M. Swapna , " Role of fuzzy logic in Data Mining", International Journal of Advance Research in Computer Science and Management Studies Volume 2, Issue 12, December 2014 pp. 87-91)
[23] M.J Lesot, L. Mouillet and B. Bouchon-Meunier Descriptive Concept extraction with exceptions by hybrid clustering.
[24] M. Rifqi, Constructing prototypes from large databases, In Proc. of IPMU'96,1996
[25] L.A Zadeh, A note on prototype theory and fuzzy sets cognition, 12:291-297, 1982
[26] M. Friedman and M. Ming and A. Kandel. "On the theory of typicality", International Journal of Uncertainity, Fuzziness and Knowledge-Based Systems, 3(2):127-142, 1995.
[27] E. Rosch and C. Mervis , Family Resemblance: Studies of internal structure of categories, Cognitive psychology, 7:573-605, 1975).
[28] http://www.simworld.co.uk/data-mining.html
[29] https://www.analyticsvidhya.com