

Classification Techniques In Data Mining and Its Applications

Priya Viridi, Gagan Amrit Kaur

Assistant Professor

Lovely Professional University, Phagwara, Punjab

virdipriya@ymail.com, gaganamrit27@gmail.com

Abstract

Data mining is a process of finding the knowledge from large amount of data. The various techniques are available to mine useful knowledge from huge database. Data mining is utilized to find information out of data and showing it in a helpful manner to the people. This paper studied few data mining classification techniques such as classification Neural Networks, Support Vector Machine (SVM), Decision tree, Bayesian classification which are used to predict the previously unidentified class of the object. Applications of data mining techniques are also discussed in this paper.

1. Introduction

Large amount of data has generated in various fields due to information technology. This data can be used for the purpose of decision making in various fields. We can find the important information from such data by applying data mining techniques. Data mining is used to filter out knowledge from large amount of data so it is also known as Knowledge Discovery from data (KDD).[1] To extract the knowledge from data and representing it in human readable form is known as knowledge discovery process.[2] There are several significant steps which required in knowledge discovery process.

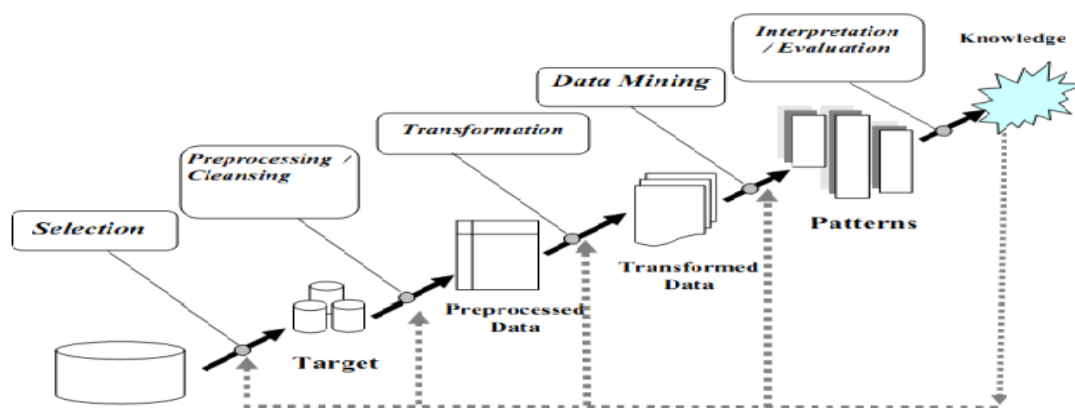


Fig 1: Knowledge Discovery Process

- **Data selection-** Relevant data for the analysis is obtained from the database.
- **Data cleaning-** To remove the erroneous and inconsistent data, data cleaning is performed
- **Data Transformation-** After data selection data is consolidated or converted in the form which is suitable for data mining.
- **Data Mining-** Data mining techniques are applied on pre-processed data in order to attain the data pattern.

- **Pattern Evaluation and Knowledge Representation-** Data Patterns are evaluated to get knowledge and knowledge representation procedures are utilized to introduce the knowledge to client.

There are various techniques are available in data mining such as Classification, clustering, regression, Association rules, Artificial intelligence. But classification and clustering are most commonly used techniques for data mining.

2. Classification

Classification is a data mining technique that assigns objects in a group to target classes. The purpose of classification is to exactly calculate the target class for each object in data.[3] It is a two step process. The initial step is learning in which classification algorithm is analyzed the training data. The subsequent step is a classification wherein test information is utilized to compute the precision of data.[4] Classification predicts the outcome dependent on determined data. Item is place with which class is determined by grouping calculations dependent on the training dataset. There are various classification techniques that can be applied on dataset.

3. Classification Techniques

- Neural Networks
- Support Vector Machines (SVM)
- Classification by decision tree induction
- Bayesian Classification

3.1 Neural Networks

The fundamental thought Neural Network is to simulate heap of thickly interconnected cells inside a computer to formulate a decision like people.[5][6] We don't have to program it to learn explicitly, it learns by itself just like a brain. Human brain contains billions of neurons. Each neuron is made up of cell body with number of connections coming off it, numerous dendrites and single axon. Dendrites are the cell's inputs carrying information towards the cell. Axon is cell's output carrying information away. Network has multiple non-natural neurons called units. Units ordered in a series of layers- Input layer, output layer, hidden layer. Input layer accepts the information from the outside, and processing of that input is done by the hidden layer. The hidden layer plays an important role in producing output. The output layer is used provide the output to the user.[7]

Input layer - receives a various form of information from the outside.

Hidden layer – actual processing is done by hidden layer

Output layer-provides the output.

A Neuron has several inputs and one output as shown in Figure 2. It works in two different modes- training mode and using mode. In training mode, the neuron is taught to activate for a specific pattern and in using mode, when a given taught pattern is encountered at the input, its affiliate output becomes existent output. Each input is affiliate with a weight. A number multiplied with the input is known as Input weight. The addition is applied to the weighted inputs and if the sum exceeds the predetermined threshold value then neuron fires for the specific input pattern. Mathematically it can be represented as:

$$I_1W_1 + I_2W_2 + I_3W_3 > T$$

Here I stand for input, W stands for weight and T stands for the threshold.

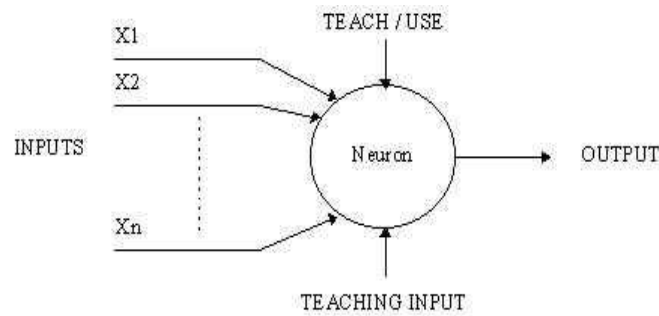


Fig 2: A simple neuron

3.2 SVM(Support Vector Machine)

SVM (support vector machine) is supervised machine learning technique used to classify the data. SVM provides higher accuracy than the other classification techniques. The basic idea behind SVM is to maximize the margin of data by discovering the best possible separating hyper plane. In the figure3 there are multiple hyper planes which separate the data but best one is chosen[8][9].

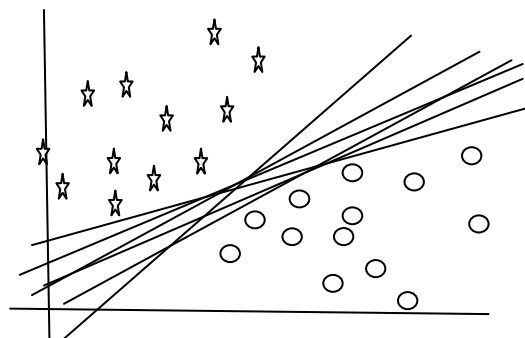


Fig 3: Multiple separating hyper planes

After that to get the margin, the difference between hyper plane and the closest data point is computed and double this value. There are multiple hyper planes in Fig 3. But choose the best hyper plane which helps to classify the data accurately. Given a particular hyper plane, we can compute the distance between the hyper plane and closest data point. Once we have this value, if we double it we will get margin. There will never be any data point inside the margin as shown in Figure 4. SVM can work on both linear and non-linear data.

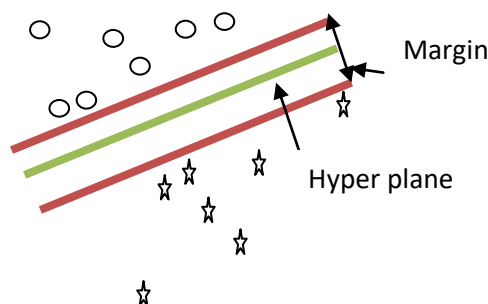


Fig 4: Optimal hyper plane and margin

3.3 Decision tree induction

A Decision tree is a flowchart-like tree structure, where each inner node means a test on a property, each branch tells to a result of the test, and each leaf node holds a class mark.[1] The highest node in a tree is the root node, it signifies the decision tree, it predicts whether it will rain or not.

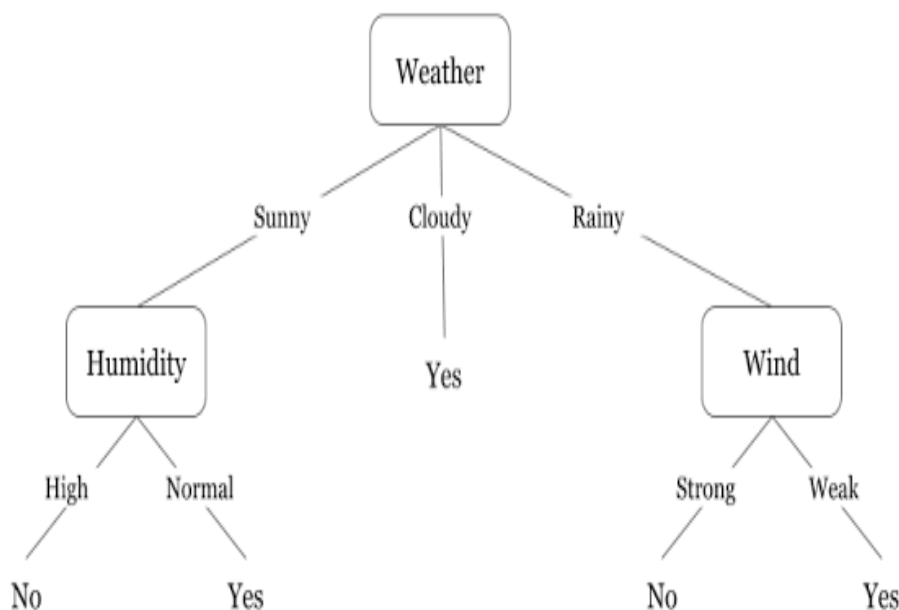


Fig 5: Decision tree for weather Forecast

The fundamental algorithm for decision tree induction is an greedy algorithm that develops decision trees in a top-bottom iterative subdivide way. The algorithm, summarized as follows.

1. make a node N;
2. if every objects belongs to similar class, C then
3. Return N as a leaf node categorized with the class C;
4. if list of feature is vacant then
5. return N as a leaf node labelled for the most general class;
6. select test-features, the feature from feature-list with the maximum information
7. test-features are labelled with N node;
8. for every identified assessment a_i of test-feature, add a branch from node N for the situation test-feature = a_i ;

3.4 Bayesian Classification

Bayesian classifiers are measurable classifiers. They can foresee class participation probabilities, for example, the likelihood that a given tuple has a place to a specific class. Bayesian grouping depends on Bayes theorem. A naive Bayes accept that the presence or absence of a specific component is random to the nearness or nonappearance of some other

element, provided the class variable. Bayesian networks, otherwise called conviction networks, have a place with the family of probabilistic graphical models[10]. These graphical structures are utilized to speak to information about a questionable area. Specifically, every hub in the chart speaks to an irregular variable, while the edges between the hubs speak to probabilistic conditions among the comparing irregular factors. These contingent conditions in the diagram are frequently assessed by utilizing known factual and computational techniques. Consequently, BNs consolidate standards from diagram theory, probability theory, computer science, and insights[11][12]

4. Data Mining Applications

Data mining application are to a great extent use in disparate territory, for example, bargain stores, medical clinics, banks, and disaster protection organizations. Numerous spaces like wellness care, financial aspects protection, and exchange necessities consolidate the data mining application with data, design appreciation and other significant devices to do data investigation. Data mining is utilized first and driving for basic leadership.

- **Medicinal and fitness concern:** Information expulsion in tablets empowers to recognize tolerant execution to see internal bound working environment visits. Data evacuation recognizes the examples of charming helpful treatments for irregular ailments.
- **Education and Training:** Education information evacuation is a maximum out earth which gives data from enlightening condition information. The objective of EDM is vital as anticipating understudy knowledge presentation, feeling and aptitudes. This learns improve the adapting strategies by quiet the swim and to get right choices similarly.
- **Research study:** Data mining is very positive in information pre-handling and amalgamation of database. Information withdrawal enables the scientists to recognize co-happening rope and the relationship flank by any introduction.
- **Farming:** Data mining is developing information in agribusiness field for produce give route psychotherapy regarding four strictures obviously year, precipitation, make and region of sows.
- **Fraud detection:** The training misrepresentation gratefulness techniques are arranged, time troubling and multipart. Information hauling out guide in given that generous examples and spinning information into in arrangement. Appropriate and obliging being developed is called as in arrangement.
- **Transportation:** Data mining help settle on a choice the conveyance plans among stockroom and outlet as sparkling as examine load designs.
- **Financial and Banking:** Data mining can make a spending to settle attention inconveniences in repository and business by decision test, causalities, and connection in huge business in assortment and in swarm costs.[13]

Conclusion

Data mining offers promising approaches to reveal shrouded designs inside a lot of data. These shrouded examples can conceivably be utilized to foresee future conduct. The accessibility of new data mining calculations, in any case, ought to be met with alert. In this paper, we present the essential classification systems. A few significant sorts of classification

strategy including Decision tree classification, Bayesian Classification, Neural Network, and Support Vector Machine. There are various applications of data mining classification techniques almost in every field. Some of the applications are also discussed in this paper.

References

- [1] J. Melton, S. Buxton, H. Samet, T. J. Teorey, S. S. Lightstone, T. P. Nadeau, J. Celko, G. Ralf, M. Schneider, J. Celko, E. Cox, T. Halpin, K. Evans, P. Hallock, B. Maclean, J. Melton, J. Melton, A. R. Simon, and M. Chisholm, "Data Mining : Concepts and Techniques". 1999.
- [2] V. P. Muthukumar, S. Subbaiah, and S. Srinivasan, "A Reveiw on Data Mining Techniques , Algorithms and tools," *IJRSC*, vol. 07, no. 01, pp. 177–190, 2018.
- [3] V. Singh, "A Survey on Classification Techniques in Data Mining," *IJRITCC*, vol. 2, no. 7, 2017.
- [4] C. Science and S. Engineering, "Research on Data Mining Classification," vol. 4, no. 4, pp. 329–332, 2014.
- [5] M. S. B. Maind, "Research Paper on Basic of Artificial Neural Network," *IJRITCC*, vol. 2, no. January, pp. 96–100, 2014.
- [6] S. Walczak, "Artificial Neural Networks," pp. 631–645, 2019.
- [7] R. Sathya and A. Abraham, "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification," *IJARI*, vol. 2, no. 2, pp. 34–38, 2013.
- [8] D. K. Srivastava and L. Bhambhu, "Data Classification Using Support Vector Machine," *JATIT*, 2009.
- [9] C. Hsu, C. Chang, and C. Lin, "A Practical Guide to Support Vector Classification," vol. 1, no. 1, pp. 1–16, 2010.
- [10] A. Bharathi and E. Deepankumar, "Survey on Classification Techniques in Data Mining," no. July, pp. 1983–1986, 2014.
- [11] B. Networks, F. Faltin, and R. Kenett, "Bayesian Networks," 2007.
- [12] I. B. Networks, "Introducing Bayesian Networks 2.1."
- [13] M. Ramageri, "Data Mining Techniques and Applications," *IJCSE*, vol. 1, no. 4, pp. 301–305.