ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 9, September 2018

A Survey Paper on Data Mining Techniques

Prof. S.P Vanjari[1], Rutuja Burate[2], Gayatri Keskar[2], Priya Dhanure[2], Sadanand Gadve[2]

Assistant Professor, Department of Information Technology, Zeal College of Engineering and Research, Narhe, Pune, Maharashtra, India[1]

B. E Student, Department of Information Technology, Zeal College of Engineering and Research, Narhe, Pune, Maharashtra, India[2]

ABSTRACT: Classification is a data mining (machine learning) technique used to predict group membership for data instances. In this paper, we present the basic classification techniques. Several major kinds of classification method including support vector machine, K-MEANS techniques. The goal of this survey is to provide a comprehensive review of different classification techniques in data mining & analyse the performance of both.
.

KEYWORDS: Support Vector Machine, Classification techniques.

I. INTRODUCTION

There are several applications for machine learning (ML), the most significant is data mining. Data mining involves sophisticated data analysis tools for discovering historical unknown, right patterns and relationships on to the large data set. These tools include statistical models, mathematical algorithm and machine learning methods. Frequently data mining consists of analysis and prediction. Classification technique applicable for processing a wider variety of data than regression. The actual data mining task is automatic analysis of large amount of data to extract historical unknown, gripping patterns such as groups of data records (cluster analysis), unusual records and dependencies. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

II. OVERVIEW OF CLASSIFICATION & CLUSTERING

Clustering & Classification are two different concept of machine learning.

Classification is one of the data mining skills that classify unstructured data into the structured class and groups and it helps to user for knowledge and finding future plan. Classification provides perspective decision making. There are two phases in classification, first is learning process phase in which a large training data sets are contributed and analysis takes place then rules and designs are created. Then the implementation of second phase begins that is evaluation or test of data sets and logs the file accurately of a classification patterns. This section briefly describes the supervised classification methods such as Support Vector Machine (SVM). Clustering is the task of combining a set of objects in such a way that objects in the same group (called a cluster) are more alike to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

DOI: 10.15680/IJIRCCE.2018. 0609028

Table1. CLASSIFICATION VS CLUSTERING

Criteria	Classification	Clustering
Knowledge of classes	Yes	No
Scenario	Classify new samples into known classes.	Suggest groups based on patterns.
Algorithms	SVM	K-MEANS
Data needs	Labelled data	Unlabelled data

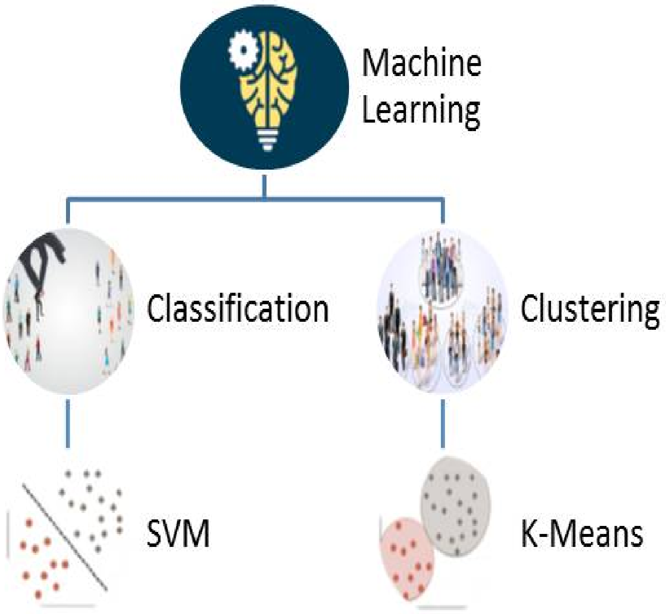

Fig 1.Machine Learning Techniques

IV. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) states as it is a discriminative classifier which is formally defined by a separating hyper plane. We can say that the given labeled training data (supervised learning).

SVM is a fast and dependable classification algorithm which performs best with a small and limited amount of data. The idea behind the SVM algorithm is simple to study and depict, and applying it to its natural language classification which doesn’t get most of the complex stuff. And that’s the basics of Support Vector Machines. A SVM allows us to classify data that is actually linearly separable. If its not then we can use the kernel view of point to make it work.

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. Examples are assigning a given email to the "spam" or "non-spam" class, and assigning a diagnosis to a given patient based on observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition.

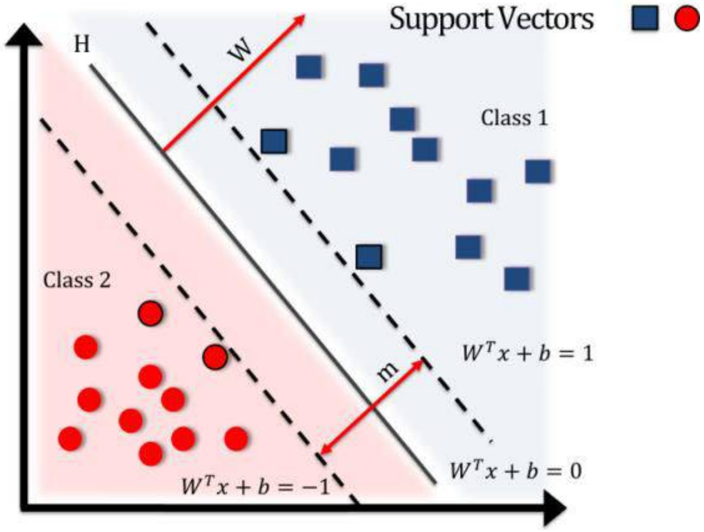

Fig 2.SVM Classification

V. K-MEANS

K-means clustering is known for unsupervised learning, which has been used when there is unlabelled data. Main motive of this algorithm just to find groups in the data, with the finding out the number of groups which is represented by the variable K. This algorithm just works step by step iteratively to give each data point one of K groups which is based on the features so to provide it. The centroids for this K clusters, which is used to label new data. Labels for this training data (each data point is assigned to a single cluster).
This algorithm, summarized as follows.
1. Initialization: The first thing k-means does, is randomly choose K examples (data points) from the dataset as initial centroids and that's simply because it does not know yet where the center of each cluster is.
2. Cluster Assignment: Then, all the data points that are the closest (similar) to a centroid will create a cluster. If we're using the Euclidean distance between data points and every centroid, a straight line is drawn between two centroids, then a perpendicular bisector (boundary line) divides this line into two clusters.

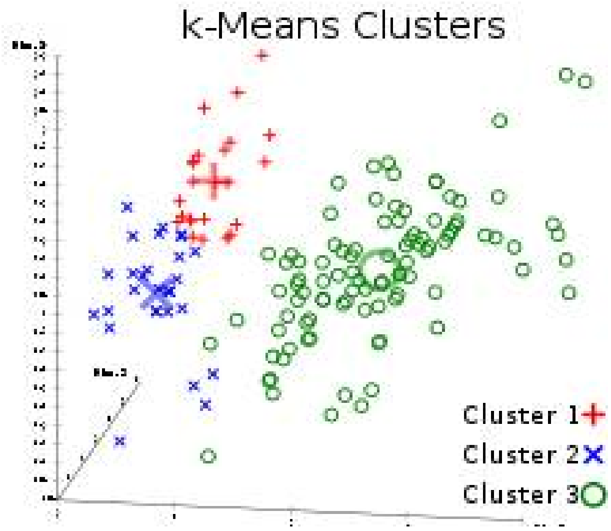

Fig 3. K-MEANS Clustering

3. Then, all the data points that are the closest (similar) to a centroid will create a cluster. If we're using the Euclidean distance between data points and every centroid, a straight line is drawn between two centroids, then a perpendicular bisector (boundary line) divides this line into two clusters.
4. Move the centroid.
5. Now, we have new clusters, that need centers. A centroid's new value is going to be the mean of all the examples in a cluster.

Advantages	Ability to learn dimensionality of the feature space.	Easy to implement With a large number of variables, K-Means may be faster than hierarchical clustering (if K is small). k-Means may produce Higher clusters than hierarchical clustering
Disadvantages	1)Kernel selection 2)Parameter tuning	1)Difficult to predict K-Value. 2)With global cluster, it didn't work well. 3) Different initial partitions can result in different final clusters.

Applications	1) Face detection 2) Text and hypertext categorization 3) Classification of images.	1)K-means clustering is rather easy to implement and apply even on large data sets, particularly when using heuristics such as Lloyd's algorithm . 2)It has been successfully used in various topics, including market segmentation, computer vision, astronomy and agriculture.

Table2. : Advantages & limitations of SVM &K-Means

VI. CONCLUSION

For this work, data mining procedure was used to extract patterns which can be profitable in predicting success and failure. The data mining techniques were applied to a work done. The data went through the cleaning and combining processes .Data mining deals with searching trends and patterns in a given data. Data mining approach is valuable since it can be easier to identify the hidden designs and relationships among various variables. These relationships can in turn help in recognizing sequence of events, classification, clustering, and predicting future events. Data mining techniques could be used in countless scenarios.

ACKNOWLEDGMENT

This work is supported in a movie prediction system of any state in India. Authors are thankful to Faculty of Engineering and Technology (FET), Savitribai Phule Pune University, Pune for providing the facility to carry out the research work.

REFERENCES

1. AGGARWAL, C.C., HINNEBURG, A., and KEIM, D.A. 2000. "**On the surprising behavior of distance metrics in high dimensional space**". IBM Research report, RC 21739. AGGARWAL, C.C.
2. PROCOPIUC, C., WOLF, J.L., YU, P.S., and PARK, J.S. 1999a. **Fast algorithms for projected clustering. In Proceedings of the ACM SIGMOD Conference**, 61-72, Philadelphia, PA.
3. AGGARWAL, C.C., WOLF, J.L., and YU, P.S. 1999b. **A new method for similarity indexing of market basket data**. In Proceedings of the ACM SIGMOD Conference, 407- 418, Philadelphia, PA.
4. AGGARWAL, C.C. and YU, P.S. 2000. **Finding generalized projected clusters in high dimensional spaces**. Sigmod Record, 29, 2, 70-92. AGRAWAL, R., FALOUTSOS, C., and SWAMI, A. 1993. Efficient similarity search in sequence databases. In Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms, Chicago, IL.
5. Wang, H., & Zhang, **"Movie genre preference prediction using machine learning for customer-based information."** 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). doi:10.1109/ccwc.2018.8301647
6. Apala, K. R., Jose, M., Motnam, S., Chan, C.-C., Liszka, K. J., & de Gregorio, F. **"Prediction of movies box office performance using social media."** Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis andMiningASONAM'13.doi:10.1145/2492517.2500232
7. Pires Dias, J. M., Oliveira, C. M., & da Silva Cruz, L. A. **" Retinal image quality assessment using generic image quality indicators."** 2014 Information Fusion, 19, 73–90. doi:10.1016/j.inffus.2012.08.00

8. Mestyán, M., Yasseri, T., & Kertész, J. (2013). **"Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data."** PLoS ONE, 8(8), e71226. doi:10.1371/journal.pone.0071226
9. Cristian E. Briguez,Maximiliano C.D. Budán, Cristhian A.D. Deagustini , Ana G. Maguitman,Marcela Capobianco , Guillermo R. Simari**"Argument-based mixed recommenders and their application to movie suggestion"** Bahía Blanca, Buenos Aires, Argentina. Published by Elsevier Ltd.2014
10. Nick Armstrong, Kevin Yoon **"Movie Rating Prediction"** Robotics Institute Carnegie Mellon University 2015
11. Andre Luiz Vizine Pereira , Eduardo Raul Hruschka **"Simultaneous co-clustering and learning to address the cold start problem in recommender systems"** University of São Paulo, São Carlos, Brazil Rubens Lara College of Technology, FATEC-RL, Santos, Brazil Published by Elsevier Ltd. 2015
12. He, G., & Lee, S. **"Multi-model or Single Model? A Study of Movie Box-Office Revenue Prediction."** 2015 IEEE International Conference on Computer and Information Technology; doi:10.1109/cit/iucc/dasc/picom.2015