IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 9, September 2016

A Survey on Data Mining using Machine Learning Techniques

Kalpana Kushwaha[1], Pinkeshwar Mishra[2]

Swami Vivekanand College of Science & Technology[1, 2]

Abstract: Data has been growing at a very fast rate. To quickly analyze this data is a challenge. This data is very useful to make decisions and predictions. So data mining for this data is very crucial. In this paper a survey is done on machine learning techniques used for data mining.

Keywords: Data Mining, Machine Learning Techniques, Data preprocessing, Data mining.

INTRODUCTION

Data is a very important asset of any organization. Company has to store all its transaction related data for its future use in business. The digital revolution provided relatively inexpensive data storage devices, which have helped the organization to store all related transaction data in the form of large information systems. Now a day because of internet usage the way transaction taking place within the organizations has completely changed. At a click of a button we can transfer the data from one part of the world into another part. Internet opened lot of opportunities for the organization to do business. Increased business opportunities create more number of possible transactions and volume of data growth. Databases today can range in size into the terabytes, more than 1, 000, 000, 000, 000 bytes of data. With in the masses of data lies hidden information of strategic importance. The quantity of data in the world roughly doubles every year[1] Tremendous data growth in the organizational databases gives the major difficulty to retrieve the hidden and useful information, which may be used for decision-making. We need a unique technique, which will work effectively to retrieve the hidden and useful decision-making information even in the midst of data growth in the databases.

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. To mine the hidden and useful information we have to take the available dataset through the process of data mining. It's not a single step. It contains various groups of interlinking steps which will help us to find the useful information for decision making. Data mining searches databases to find hidden patterns and predict information to increase the business in the organization.

Data mining Life Cycle

We have to do the following steps to solve a data mining.

Define the problem: To have the successful data mining application, the organization has to come up with a precise formulation of the problem they are trying to solve. A focused problem statement usually results in the best payoff.

Data collection and selection: The organization has to use the right data for mining. data collection and selection step identifies the related data sources and acquires it. From the collected data source data selection process selects the subset of data to mine.

Data preprocessing:

Data cleaning It fills in the missing data and correcting the invalid data into a valid one. It finds the outliers data and removes the inconsistencies in the data source.

Data integration: It combines data from different data sources into a single mining database.

Data transformation: It converts the source data into a common format for processing.

Data reduction: It is a process of discarding unwanted parameters from the data. So that the data volume will be less at the same time it will not suffer on the quality of the information.

Data discretization: It is a part of data reduction process. It replaces the numerical attributes with the nominal attributes.

LITERATURE REVIEW

One of the approaches followed to predict customer behavior is the use of the transactional data. For example, [21] developed a model using hierarchical clustering and a hidden Markov model (HMM) to predict customer behavior based on transactional data. [22] also used Markov model to predict the probability of click to conversion based on the time spent by the customer on site. Once a retailer knows the underlying behavior of a consumer, then based on the products that a customer selected in the past, they can design recommender systems

DOI 10.17148/IJARCCE.2016.5940

to assist them in selecting similar products [23]. The underlying assumption is that the consumers follow patterns similar to their past spending habits and are likely to repeat it in the future. Using different machine learning techniques such as classification, genetic algorithms, clustering or K-nearest neighbor algorithms [23], retailers can potentially identify different customer segments and predict customers' preference and spending abilities. This can help retailers in better advertising of their products to the right audiences. A number of data mining techniques to detect cyber crimes are proposed in the literature [24].

For example, classification models such as Naive Bayes, support vector machines, neural networks, decision trees have long been used to detect spam emails [25], (spamming implies sending unsolicited emails). Support vector machine techniques have also been used to prevent Denial of Service (DoS) attacks, where DoS attack refers to the process of making system inaccessible to other users [26], [27]. While [26] used Enhanced Multi Class Support Vector Machines (EMCSVM) to predict various kinds of DoS attacks, [27] proposed radial-basis function neural network (RBFNN) and support vector machines (SVM), to solve the DoS problem with an ability to detect or predict new attacks based on the patterns similar to the attack patterns that appeared in the past. Classification models have also been used to detect Malware [28] and phishing URLs [29] and emails [30].

Data mining through machine learning

To efficiently analyze this big data machine learning techniques are very important. This data may be structured and unstructured. Structured data consists of large number of features in case of Big data. So for accurate analysis of such data features also need to be reduced such analysis is termed as principal component analysis.

Supervised Learning techniques

K-Nearest Neighbors: In pattern recognition, the k-Nearest Neighbors algorithm (or k-NN for short) is a non-parametric method used for classification and regression.[1] In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression: In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors. k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

Naive Bayes: It is a probability based classification technique. It considers all features independent of each other. It calculates probability of each feature independently for a particular class label. Mathematically it can be denoted as:

P(x/y) which denotes probability of feature x in the feature set given a class label 'y'. Then for all the features total probability will be:

$$P(x/y) = \prod_{k=1}^{d} p\left(\frac{x_k}{y}\right)$$

Then the posterior probability of class 'y' given that x feature is in the feature set is given by:

$$P\left(\frac{y}{x}\right) = \frac{P\left(\frac{x}{y}\right)P(y)}{P(x)} = \frac{P\left(\frac{x}{y}\right)P(y)}{\sum_j P\left(\frac{x}{y_j}\right)P(y_j)}$$

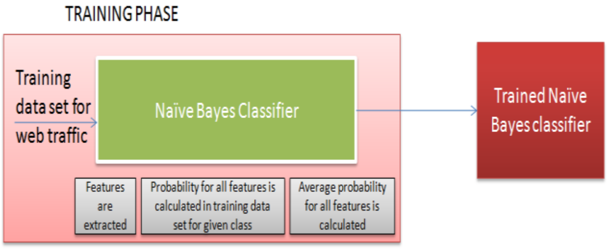
The features for which P(y/x) is more are the most deciding features and can also be considered as principle components.

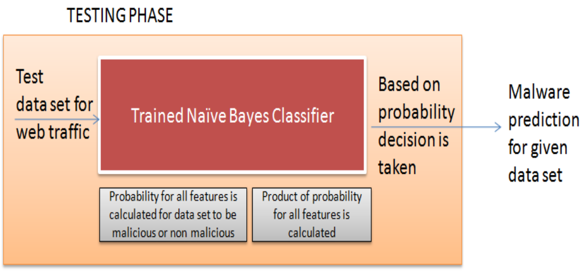
Advantages:

a. Since this approach is based on the probability it can be applied to a wide variety of domains and results can be used in many ways.
b. Doesn't require large amount of data before training to begin.
c. These algorithms are computationally fast to make decisions.

Naïve Bayes is used in this paper for malware prediction using web traffic data. These are the steps behind the Naïve Bayes algorithm:

1. Training data set is taken as input.
2. Features are extracted from that training data. In this paper web traffic data consists of 43 features.
3. Then from the training data for every feature Naïve bayes calculates probability that if feature has particular value then the dataset class be will malicious or not.
4. If every feature has limited possible values then above probabilities can be calculated. But if the large number of values is there for every feature, range of values can also be taken.
5. Then for every row of test data set after the training phase. On the basis of average probabilities calculated from training data decision is taken.


TRAINING PHASE
Training data set for web traffic
Naïve Bayes Classifier
Trained Naïve Bayes classifier
Features are extracted
Probability for all features is calculated in training data set for given class
Average probability for all features is calculated


TESTING PHASE
Test data set for web traffic
Trained Naïve Bayes Classifier
Based on probability decision is taken
Malware prediction for given data set
Probability for all features is calculated for data set to be malicious or non malicious
Product of probability for all features is calculated

Support Vector Machine: It is a classifier which finds a hyper plane that clearly separates the sample points of different labels. And it divides such that sample points of both labels or class are on different sides of hyper plane. The hyper plane is generated such that it satisfies two constraints:

a. It should separate sample point of both labels.
b. Distance of closest sample point of both labels should be maximum.

Mathematically hyper plane is denoted as:

$$w.x - b = 0$$

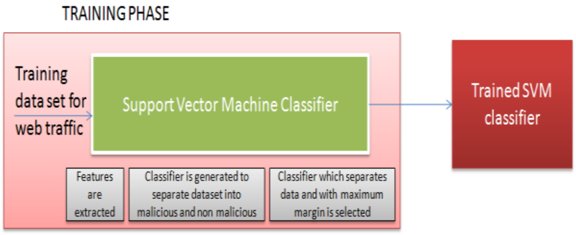
where, . denotes the dot product,
w is normal vector, and the parameter b/||w|| determines the offset of the hyper plane from the origin along the normal vector w.
W and b should be chosen such that margin should be maximum and distance between parallel hyper planes should be maximum and should still separate the sample points of labels given.

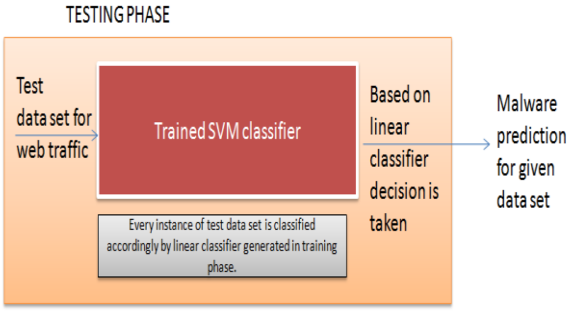
Disadvantages:

a. Biggest limitation of SVM is appropriate selection of kernel according to the dataset.
b. Second is speed is slow and gets even slower with size of testing and training dataset.

SVM can also be used for web based malware prediction using traffic data. It involves following steps:

1. Training data set is taken as input.
2. Features are extracted from that training data.
3. Classifier is generated which separates the data into malicious and non malicious data.
4. The best classifier is the one which has maximum margin and successfully separates the 2 classes.
5. Test data is given and every instance of the data is classified according to the generated classifier.


TRAINING PHASE
Training data set for web traffic
Support Vector Machine Classifier
Trained SVM classifier
Features are extracted
Classifier is generated to separate dataset into malicious and non malicious
Classifier which separates data and with maximum margin is selected


TESTING PHASE
Test data set for web traffic
Trained SVM classifier
Based on linear classifier decision is taken
Malware prediction for given data set
Every instance of test data set is classified accordingly by linear classifier generated in training phase.

Decision tree: This type of classifier models data with the help of a tree. Tree is having features as the internal nodes and edges indicate the values of features. And edges separated nodes based on the values. All the leaf nodes of the decision tree represents a class which is expected to be obtained if we have all the features having respective values which are in the path from the root to that class having intermediate feature nodes.

Some of the most popular decision tree algorithms are ID3, C4.5, CART. ID3 is one of the simplest decision tree approaches it uses concept of information gain as the splitting criteria. C4.5 is the evolution of ID3. It works on the principle of gain ratio.

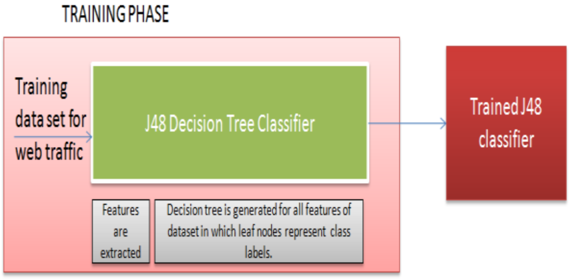
Advantages: All decision tree approaches are simple to understand and easy to interpret.

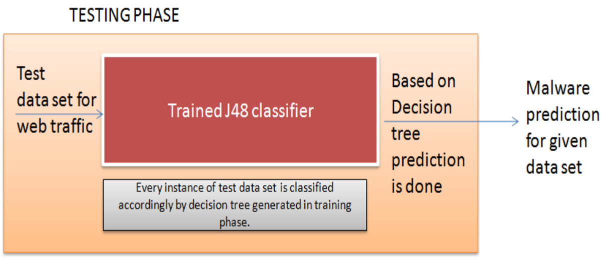
Disadvantages:

a. Most of the decision tree algorithms require features to have only discrete values.
b. As these algorithms work on the principle of divide and conquer method so these algorithms perform well only if attributes are highly relevant to each other. But poor when complex relationships are their in feature set.

Decision tree can be used to detect malwares using following steps:

1. Training data set is taken as input.
2. Features are extracted from that training data.
3. Decision tree is generated based on the relation between the features such that leaf nodes of tree represent class labels.
4. Test data is given and every instance of the data is classified according to the generated decision tree.


TRAINING PHASE
Training data set for web traffic
J48 Decision Tree Classifier
Trained J48 classifier
Features are extracted
Decision tree is generated for all features of dataset in which leaf nodes represent class labels.


TESTING PHASE
Test data set for web traffic
Trained J48 classifier
Every instance of test data set is classified accordingly by decision tree generated in training phase.
Based on Decision tree prediction is done
Malware prediction for given data set

REFERENCES

[1] Xin Bi, Xiangguo Zhao, Guoren Wang, Pan Zhang, and Chao Wang. Distributed extreme learning machine with kernels based on MapReduce. Neurocomputing, 149:456–463, 2015.

[2] Jiaoyan Chen, Guozhou Zheng, and Huajun Chen. ELM-MapReduce: MapReduce accelerated extreme learning machine for big spatial data analysis. In 10th International Conference on Control and Automation (ICCA), pages 400–405. IEEE, 2013.

[3] George Cybenko. Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4):303–314, 1989.

[4] Qing He, Tianfeng Shang, Fuzhen Zhuang, and Zhongzhi Shi. Parallel extreme learning machine for regression based on MapReduce. Neurocomputing, 102:52–58, 2013.

[5] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. Neural networks, 4(2):251–257, 1991.

[6] G Huang, S Song, JN Gupta, and C Wu. Semi-supervised and unsupervised extreme learning machines. IEEE transactions on cybernetics, 44(12):2405–2417, 2014.

[7] Gao Huang, Guang Bin Huang, Shiji Song, and Keyou You. Trends in extreme learning machines: A review. Neural Networks, 61:32–48, 2015.

[8] Guang Bin Huang, Lei Chen, and Chee Kheong Siew. Universal approximation using incremental constructive feedforward networks with random hidden nodes. Transactions on Neural Networks, 17(4):879–892, 2006.

[9] Guang Bin Huang, Qin Yu Zhu, and Chee Kheong Siew. Extreme learning machine: a new learning scheme of feedforward neural networks. In Proceedings of the International Joint Conference on Neural Networks, volume 2, pages 985–990. IEEE, 2004.

[10] M. Lichman. UCI machine learning repository, 2013.

[11] Andrew McCallum, Kamal Nigam, and Lyle H Ungar. Efficient clustering of high-dimensional data sets with application to reference matching. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 169–178. ACM, 2000.

[12] Mark Van Heeswijk, Yoan Miche, Erkki Oja, and Amaury Lendasse. GPU-accelerated and parallelized ELM ensembles for large-scale regression. Neurocomputing, 74(16):2430–2437, 2011.

[13] Botao Wang, Shan Huang, Junhao Qiu, Yu Liu, and Guoren Wang. Parallel online sequential extreme learning machine based on MapReduce. Neurocomputing, 149:224–232, 2015.

[14] Ran Wang, Yu Lin He, Chi Yin Chow, Fang Fang Ou, and Jian Zhang. Learning ELM-tree from big data based on uncertainty reduction. Fuzzy Sets and Systems, 258:79–100, 2015.

[15] Junchang Xin, Zhiqiong Wang, Chen Chen, Linlin Ding, Guoren Wang, and Yuhai Zhao. ELM*: distributed extreme learning machine with MapReduce. World Wide Web, 17(5):1189–1204, 2014.

[16] Junhai Zhai, Jinggeng Wang, and Xizhao Wang. Ensemble online sequential extreme learning machine for large data set classification. In International Conference on Systems, Man and Cybernetics, pages 2250–2255. IEEE, 2014.

[17] H Zhou, GB Huang, Z Lin, H Wang, and YC Soh. Stacked extreme learning machines. IEEE transactions on cybernetics, PP(99):1–13, 2014.

[18] Saeed Shahrivari, "Beyond Batch Processing: Towards Real-Time and Streaming Big Data", Computers, Vol. 3, pp. 117.129, 2014.

[19] Sachin Agarwal. Monitoring and Troubleshooting Apache Storm [online]. Available: https://dzone.com/articles/monitoring-and-troubleshooting-apache-storm-with-o. Date accessed: (April 7, 2016).

[20] M. R. Evans, D. Oliver, K. Yang, X. Zhou, S. Shekhar, "Enabling Spatial Big Data via CyberGIS: Challenges and Opportunities," Ed. S. Wang, M. F. Goodchild, CyberGIS: Fostering a new wave of Geospatial Innovation and Discovery. Springer, 2014.

[21] M. Mestre and P. Vitoria, "Tracking of consumer behaviour in ecommerce," in Information Fusion, 16th International Conference on, July 2013, pp. 1214–1221.

[22] M. Gupta, H. Mittal, P. Singla, and A. Bagchi, "Characterizing comparison shopping behavior: A case study," in Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on, March 2014, pp. 115–122.

[23] H.-W. Yang, Z. geng Pan, X.-Z. Wang, and B. Xu, "A personalized products selection assistance based on e-commerce machine learning," in Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on, vol. 4, Aug. 2004, pp. 2629–2633.

[24] T. Mahmood and U. Afzal, "Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools," in Information Assurance (NCIA), 2013 2nd National Conference on, Dec 2013, pp. 129–134.

[25] P. Panigrahi, "A comparative study of supervised machine learning techniques for spam e-mail filtering," in Computational Intelligence and Communication Networks, Fourth International Conference on, Nov 2012, pp. 506–512.

[26] T. Subbulakshmi, S. Shalinie, V. GanapathiSubramanian, K. BalaKrishnan, D. AnandKumar, and K. Kannathal, "Detection of ddos attacks using enhanced support vector machines with real time generated dataset," in Advanced Computing (ICoAC), 2011 Third International Conference on, Dec 2011, pp. 17–22.

[27] G. Tsang, P. Chan, D. Yeung, and E. Tsang, "Denial of service detection by support vector machines and radial-basis function neural network," in Machine Learning and Cybernetics, Proceedings of 2004 International Conference on, vol. 7, Aug 2004, pp. 4263–4268.

[28] M. Mas'ud, S. Sahib, M. Abdollah, S. Selamat, and R. Yusof, "Analysis of features selection and machine learning classifier in android malware detection," in Information Science and Applications, International Conference on, May 2014, pp. 1–5.

[29] J. James, L. Sandhya, and C. Thomas, "Detection of phishing urls using machine learning techniques," in Control Communication and Computing, International Conference on, Dec 2013, pp. 304–309.

[30] A. Almomani, B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani, "A survey of phishing email filtering techniques," Communications Surveys Tutorials, IEEE, vol. 15, no. 4, pp. 2070–2090, Fourth 2013.

[31] B. Thuraisingham, "Data mining for security applications," in Machine Learning and Applications, 2004. Proceedings. 2004 International Conference on, Dec 2004, pp. 3–4.

[32] A. Aziz, A. Hassanien, S.-O. Hanaf, and M. Tolba, "Multi-layer hybrid machine learning techniques for anomalies detection and classification approach," in Hybrid Intelligent Systems (HIS), 2013 13th International Conference on, Dec 2013, pp. 215–220.

DOI 10.17148/IJARCCE.2016.5940