https://dx.doi.org/10.22161/ijaers/310.24 *ISSN: 2349-6495(P) | 2456-1908(O)*

A Survey of Classification Methods

Somia B. Mohammed[1], Ahmed Khalid[2], SaifeEldin F. Osman[3]

[1]College of Higher Education, Al Rebat University, Sudan
[2]Department of computer science, Najran University, KSA
[3]Computer science Department, Emirates College for Science and Technology, Sudan

Abstract*—Classification may refer to categorization, the process in which ideas and objects are recognized, differentiated, and understood. There are many types of classification, researchers face a problem to choose a suitable method that give a good classification performance to solve their classification problems. In this paper, we present the basic classification techniques. Several major kinds of classification method including neural network, decision tree, Bayesian networks, support vector machine and k-nearest neighbor classifier. The goal of this survey is to provide a comprehensive review of the above different classification techniques.*

Keywords—Classification methods Neural network linear classification Decision trees.

I. INTRODUCTION

Classification methods are the way of classifying data into predefined classes. Classification method uses a set of features or parameters to characterize each object, where these features should be relevant to the task at hand [1]. Classification, which maps a data item into one of several, predefined categories. These algorithms normally output "classifiers" has ability to classify new data in the future, for example, in the form of decision trees or rules. An ideal application in intrusion detection will be together sufficient "normal" and "abnormal" audit data for a user or a program. Here audit data refers to (pre-processed) records, each with a number of features (fields). Then a classification algorithm has been applied to train a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class [2]. Many decision-making tasks are instances of classification problem or can be easily formulated into a classification problem, e.g., prediction and forecasting tasks, diagnosis tasks, and pattern recognition [15]. The research on linear classification has been a very active topic [16]. With the increasing of Internet scale network traffic classification is more and more important in network security, traffic scheduling and traffic accounting etc. [17,18].

Classification will be used when an object needs to be classified into a predefined class or group based on attributes of that object. There are many real world applications that can be categorized as classification problems such as weather forecast, credit risk evaluation, medical diagnosis, bankruptcy prediction, speech recognition, handwritten character recognition [19] and Survival analysis [20].

In recent studies the performance of different classification techniques have been based mainly on experimental approaches [9,10, 11]. Empirical comparisons among different classification methods suggest that no single method is best for all learning classification tasks [12,13]. In other words, each method is best for some, but not for all tasks.

Classification systems play an important role in business decision-making tasks by classifying the available information based on some criteria.[4]. The objective of this paper is to reviews the well-known classification methods neural networks, decision trees, k-Nearest Neighbor, Naive Bayes, and Support Vector Machines. The rest of this paper is organize as follow: Our next section presents neural network. Section 3 describes decision tree. Naive Bayesian Network is discusses in section four. Section five gives details about support vector machine. Where k-nearest neighbor classifier is presents in section six. Finally, the last section concludes this work.

II. NEURAL NETWORK CLASSIFICATION METHODS.

Many types of Neural Networks can be used for classification but most popular NN is Back propagation NN and RBF NN. Artificial neural networks were initially developed according to the elementary principle of the operation of the (human) neural system [22]. Since then, a very large variety of networks have been constructed. All are composed of units (neurons), and connections between them, which together determine the behavior of the network.

2.1- Backpropagation Neural Network:

It is shown that from the literature review a BPNN having single layer of neurons could classify a set of points perfectly if they were linearly separable. BPNN having three layers of weights can generated arbitrary decision regions which may be non-convex and disjoint.

https://dx.doi.org/10.22161/ijaers/310.24 *ISSN: 2349-6495(P) | 2456-1908(O)*

BPNN is based on processing elements, which compute a nonlinear function of the scalar product of the input vector and a weight vector [5].

One of the most popular NN algorithms is back propagation algorithm [23].Claimed that BP algorithm could be broken down to four main steps. After choosing the weights of the network randomly, the back propagation algorithm is used to compute the necessary corrections. The algorithm can be decomposed in the following four steps:

i) Feed-forward computation
ii) ii) Back propagation to the output layer
iii) iii) Back propagation to the hidden layer
iv) iv) Weight updates

Here are some situations where a BP NN might be a useful:

- A large amount of input/output data is available, but you're not sure how to relate it to the output.
- The problem appears to have overwhelming complexity, but there is clearly a solution.
- It is easy to create a number of examples of the correct behavior.
- The solution to the problem may change over time, within the bounds of the given input and output parameters (i.e., today 2+2=4, but in the future we may find that 2+2=3.8).
- Outputs can be "fuzzy", or non-numeric.

Linear classification is a useful tool in machine learning and data mining. In contrast to nonlinear classifiers such as kernel methods, which map data to a higher dimensional space, linear classifiers directly work on data in the original input space. While linear classifiers fail to handle some inseparable data, they may be sufficient for data in a rich dimensional space. For example, linear classifiers have shown to give competitive performances on document data with nonlinear classifiers. An important advantage of linear classification is that training and testing procedures are much more efficient. Therefore, linear classification can be very useful for some large-scale applications. Recently, the research on linear classification has been a very active topic. In this paper, we give a comprehensive survey on the recent advances [6].

III. DECISION TREE

Decision tree is classification scheme which generates a tree and asset of rules representing the model of different classes, from a given dataset .As in [41], DT is a flow chart like tree structure, where each internal node denotes a test on an attribute ,each branch represents an outcome of the test and leaf node represent the classes or class distributions .the top most node in a tree is the root node.[25]

Decision trees are usually unvaried since they use based on a single feature at each internal node. Most decision tree algorithms cannot perform well with problems that require diagonal partitioning. The division 0f the instance space is orthogonal to the axis of one variable and parallel to all other axes. Therefore, the resulting regions after partitioning are all hyper rectangles. However, there are a few methods that construct multivariate trees. One example is as in [43],

Decision trees can be significantly more complex representation for some concepts due to the replication problem. A solution is using an algorithm to implement complex features at nodes in order to avoid replication,.Markovitch and Rosenstein in [42] presented the FICUS construction algorithm, which receives the standard input of supervised learning as well as a feature representation specification, and uses them to produce a set of generated features. While FICUS is similar in some aspects to other feature construction algorithms, its main strength is its generality and flexibility. FICUS was designed to perform feature generation given any feature representation specification complying with its general purpose grammar. The most well-known algorithm in the literature for building decision trees is the C4.5 (Quinlan, 1993) [44]. C4.5is an extension of Quinlan's earlier ID3 algorithm.

IV. NAIVE BAYES CLASSIFIER

Bayesian networks can efficiently represent complex probability distributions, and have received much attention in recent years[14].During the past decade Bayesian networks have gained popularity in AI as a means of representing and reasoning with uncertain knowledge. Examples of practical applications include decision support, safety and risk evaluation, control systems, and data mining [26]. In the software engineering field, Bayesian networks have been used by Fenton [46] for software quality prediction. Naive Bayes is one of the most effective and efficient classification algorithms [24]. In classification learning problems, a learner attempts to construct a classifier from a given set of training examples with class labels. Abstractly, the probability model for a classifier is a conditional model

$$P(C|F_1, \dots, F_n)$$

Over a dependent class variable C with a small number of outcomes or classes, conditional on several feature variables F1 through Fn. the problem is that if the number of features n is large or when a feature can take on a large number of values, then basing such a model on probability tables is infeasible. We therefore reformulate

https://dx.doi.org/10.22161/ijaers/310.24 *ISSN: 2349-6495(P) | 2456-1908(O)*

the model to make it tractable. Using bayes' theorem, we can write

$$P(C|F_1, \dots, F_n) = \frac{P(C)P(F_1,\dots,F_n|C)}{P(F_1,\dots,F_n)}$$

In plain English the above equation can be written as

$$posterior = \frac{prior \times likelihood}{evidence}$$

In practice we are interested in the numerator of the fraction, since the denominator does not depend on C and the value of the features Fi are given, so the denominator is effectively constant. The numerator is equivalent to the joint probability model

$$P(C|F_1, \dots, F_n)$$

Which can be rewritten as follows, using repeated application of the definition of conditional probability:

$P(C_1, F_1, \dots, F_n)$

$= P(C)P(F_1, \dots, F_n|C)$

$= P(C)(F_1|C)P(F_2, \dots, F_n|C, F_1)$

$= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3, \dots, F_n|C, F_1, F_2)$

$= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2)P(F_4, \dots, F_n|C, F_1, F_2, F_3)$

$= P(C)P(F_1|C)P(F_2|C, F_1)P(F_3|C, F_1, F_2), \dots, P(F_n|C, F_1, F_2, F_3, \dots, F_{n-1})$

Now the "Naïve" conditional independence assumptions come into play: assume that each feature F_i is conditionally independent of every other feature F_j for $j \neq i$. This means that

$$P(F_i|C, F_j) = P(F_i|C)$$

For $i \neq j$, and so the joint model can be expressed as

$P(C_1, F_1, \dots, F_n) = P(C)P(F_1|C)P(F_2|C)P(F_3|C) \dots$

$= P(C) \prod_{i=1}^{n} P(F_i|C)$

This means that under the above independent assumptions, the conditional distribution over the class variable C can be expressed like this:

$$P(C_1, F_1, \dots, F_n) = \frac{1}{Z} P(C) \prod_{i=1}^{n} P(F_i|C)$$

Where Z is scaling factor dependent only on $F_1, \dots, F_n$, i.e., a constant if the values of the feature variables as known.

Model of this form are much more manageable, since they factor into a so-called class prior P(C) and independent probability distributions P(Fi|C). If there are k classes and if a model for each P(Fi|C=c)can be expressed in terms of r parameters, then the corresponding naïve bayses model has (k-1) + nrk parameters. In practice, often k=2 and r=1 are common, and so the total number of parameters of the naïve bayes model is 2n +1, where n is the number of binary features used for classification and prediction.

General Bayesian network classifiers are known as Bayesian networks, belief networks or causal probabilistic networks. The theoretical concepts of Bayesian networks were invented by Judea Pearl in the 1980s and are described in his pioneering book Probabilistic Reasoning in Intelligent Systems [27]. During the past decade Bayesian networks have gained popularity in AI as a means of representing and reasoning with uncertain knowledge. Examples of practical applications include decision support, safety and risk evaluation, control systems, and data mining [32]. The state-of-the-art research papers on Bayesian networks are published in the proceedings of the Annual Conference on Uncertainty in AI [33]. Theoretical principles of Bayesian networks are described in several books, for example [27-31].

V. SUPPORT VECTOR MACHINES

Support Vector Machine (SVM) was first heard in 1992, introduced by Boser, Guyon and Vapnik in COLT-92. Support vector machines (SVMs) are a set of related supervised learning Methods used for classification and regression [8]. They belong to a family of generalized linear classifiers. In another terms, Support Vector Machine (SVM) is a classification and regression prediction tool that uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fit to the data. SVMs were developed to solve the classification problem, but recently they have been extended to solve regression problems [14].

VI. K-NEAREST NEIGHBOR

K-Nearest Neighbor is one of the most popular algorithms for text categorization [34]. Many researchers have found that the KNN algorithm achieves very good performance in their experiments on different data sets [35,7,36]. The idea behind k-Nearest Neighbor algorithm is quite straightforward. To classify a new document, the system finds the k nearest neighbors among the training documents, and uses the categories of the k nearest neighbors to weight the category candidates [34]. One of the drawbacks of KNN algorithm is its efficiency, as it needs to compare a test document with all samples in the training set. In addition, the performance of this algorithm greatly depends on two factors, that is, a suitable similarity function and an appropriate value for the parameter k. The KNN is the fundamental and simplest classification technique when there is little or no prior knowledge about the distribution of the data [37-40].

VII. CONCLUSIONS

This paper gives a survey of classification methods focusing on neural network, decision tree, Bayesian

https://dx.doi.org/10.22161/ijaers/310.24 *ISSN: 2349-6495(P) | 2456-1908(O)*

networks, support vector machine and k-nearest neighbor classifier.

REFERENCES

[1] Moore, Andrew W., and Denis Zuev. "Internet traffic classification using bayesian analysis techniques." *ACM SIGMETRICS Performance Evaluation Review*. Vol. 33. No. 1. ACM, 2005.

[2] Helali, Rasha G. Mohammed. "Data mining based network intrusion detection system: A survey." *Novel Algorithms and Techniques in Telecommunications and Networking*. Springer Netherlands, 2010. 501-505.

[3] Pearl, J. (1988). *Probabilistic reasoning in intelligent systems,Networks of plausible inference*. San Francisco, CA: Morgan Kaufmann

[4] Kiang, Melody Y. "A comparative assessment of classification methods." *Decision Support Systems* 35.4 (2003): 441-454.

[5] J. D. Dhande1 , Dr. S.M. Gulhane2. "Design of Classifier Using Artificial Neural Network for Patients Survival Analysis",International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 1, Issue 2, November 2012 278

[6] Guo-Xun Yuan, Chia-Hua Ho, and Chih-Jen Lin, "Recent Advances of Large-scale Linear Classification",The Journal of Machine Learning Research archive, Volume 9, 6/1/2008pp 1871-1874.

[7] Joachims T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features [A]. In: Proceedings of the European Conference on Machine Learning [C].

[8] Wikipedia Online. Http://en.wikipedia.org/wiki

[9] H. Almuallim, T.G. Dietterich, Learning Boolean concepts in the presence of many irrelevant features, Artificial Intelligence 69 (1994) 279 – 305.

[10] T.G. Dietterich, H. Hild, G. Bakiri, A comparison of ID3 and backpropagation for english text-to-speech mapping, Machine Learning 18 (1995) 51 – 80.

[11] D. Wettschereck, T.G. Dietterich, An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms, Machine Learning 19 (1995) 5 – 27.

[12] S. Salzberg, A nearest hyperrectangle learning method, Machine Learning 6 (1991) 277 – 309.

[13] J.W. Shavlik, R.J. Mooney, G.G. Towell, Symbolic and neural learning algorithms: an experimental comparison, Machine Learning 6 (1991) 111 – 144.

[14] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in NeuralInformation Processing Systems 9, pages 281–287, Cambridge, MA, 1997. MIT Press.

[15] Melody Y. Kiang , "A comparative assessment of classification methods", Decision Support Systems 35 (2003) 441 – 454.

[16] Yuan, G.-X., Ho, C.-H., Lin, C.-J.: Recent Advances of Large-Scale Linear Classification. Technical report (2012).

[17] A. Sperotto, G. Schffrath, and R. Sadre, et al., "An overview of IP flow-based intrusion detection," IEEE Communications Surveys and Tutorials, vol. 12, pp. 1-14, 2010.

[18] Z, X. SUN and J. LIN, "Research of intelligent rule-base based on multilayer intrusion detection," Journal of Computers, vol. 4, no. 6, pp. 453-460, 2009

[19] Zhange G.P, Neural Networks for classification survey; System, Man, and cybernetics, part C: Application and Reviews, IEEE Transaction on, Vol.30, pp 451-462, 2000.

[20] J. D. Dhande, Dr. S.M. Gulhane, "Design of Classifier Using Artificial Neural Network for Patients Survival Analysis", International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 1, Issue 2, November 2012 278

[21] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 2nd edition, 1998

[22] David Reby , SovanLek, IoannisDimopoulos , Jean Joachim , Jacques Lauga, Ste´phaneAulagnie "Artificial neural networks as a classification method in the behavioural sciences", Behavioural Processes 40 (1997) 35–43.

[23] D.T. Larose. Discovering knowledge in data: an introduction to data mining. Wiley-Interscience,2005.78,0471666578.,URL ,http://books.google.ie/books?id=JbPMdPWQIOwC.

[24] Zhang, Harry, and Jiang Su. "Naive bayesian classifiers for ranking." European Conference on Machine Learning. Springer Berlin Heidelberg, 2004.

[25] Han ,J, Kamber , M. Data mining concepts and techniques ,morgankaufmann publisher ,2001

[26] Hugin Expert, Aalborg, Denmark, 2001. http://www.hugin.com/cases/ [27 March 2002].

[27] Pearl J. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann: San Mateo CA, 1988.

[28] Neapolitan RE. Probabilistic Reasoning in Expert Systems: Theory and Algorithms. John Wiley: New York NY, 1990.

[29] Castillo E, Gutierrez JM, Hadi AS. Expert Systems and Probabilistic Network Models. Springer: New York NY, 1997.

https://dx.doi.org/10.22161/ijaers/310.24 *ISSN: 2349-6495(P) | 2456-1908(O)*

[30] Cowell RG, Dawid AP, Lauritzen SL, Spiegel halter DJ. Probabilistic Networks and Expert Systems. Springer: New York1NY, 1999.
[31] Jensen FV. An Introduction to Bayesian Networks. UCL Press: London, 1996.
[32] Hugin Expert, Aalborg, Denmark, 2001. http://www.hugin.com/cases/.
[33] Association for Uncertainty in Artificial Intelligence, 2001. http://www.auai.org/ .
[34] Manning C. D. and Schutze H., 1999. Foundations of Statistical Natural Language Processing [M]. Cambridge: MIT Press.
[35] Yang Y. and Liu X., 1999. A Re-examination of Text Categorization Methods [A]. In: Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. 42-49.
[36] Li Baoli, Chen Yuzhong, and Yu Shiwen, 2002. A Comparative Study on Automatic Categorization Methods for Chinese Search Engine [A]. In: Proceedings of the Eighth Joint International Computer Conference [C]. Hangzhou: Zhejiang University Press, 117-120.
[37] Devroye, L. (1981) "On the equality of Cover and Hart in nearest neighbor discrimination", IEEE Trans. Pattern Anal. Mach. lntell. 3: 75-78.
[38] Devroye, L., Gyorfi, L., Krzyzak, A. & Lugosi, G. (1994) "On the strong universal consistency of nearest neighbor regression function estimates", Ann. Statist, 22: 1371–1385.
[39] Devroye, L. & Wagner, T.J. (1977) "The strong uniform consistency of nearest neighbor density estimates", Ann. Statist., 5: 536–540.
[40] Devroye, L. & Wagner, T.J. (1982) "Nearest neighbor methods in discrimination, In Classification, Pattern Recognition and Reduction of Dimensionality", Handbook of Statistics, 2: 193–197. North-Holland, Amsterdam
[41] Han , J ,kamber ,M.Data mining concepts and Techniques morgan kaufmann publisher , 2001
[42] Brighton, H. & Mellish, C. (2002), Advances in Instance Selection for Instance-Based Learning Algorithms. Data
[43] Mining and Knowledge Discovery 6: 153–172.
[44] M. Sujatha, S. Prabhakar, Dr. G. Lavanya Devi, “A Survey of Classification Techniques in Data Mining” International Journal of Innovations in Engineering and Technology (IJIET)Vol. 2 Vol. 2 Issue 4 August 2013.
[45] Quinlan , J .R .(1993) , C 4.5: programs for Machine Leaning , Margan Kaufmann , San mated ,Califarnia.
[46] Agena, London, UK, 2001. http://www.agena.co.uk/.