© 2014 IJEDR | Volume 2, Issue 4 | ISSN: 2321-9939

A Survey on Classification Techniques in Data Mining

Dr.R.Nallusamy[1], R.Anbu selvi[2]
Principal[1], PG Scholar[2]
Nandha College of Technology, Erode

Abstract **- Data mining is a process of extracting the knowledge pattern from large data.Classification is a technique which used to predict the group membership for data sets.There are various techniques such asSupport Vector Machine, Bayesian networks and the k-nearest neighbor classifier are analyzed here.The goal of this survey is to provide an expandedreview of different classification techniques in data mining.**

Keywords **- Support Vector Machine, CART, Minimum Description length, K-Nearest Neighbor,Bayesian network andInstance Based Learning.**

I. INTRODUCTION

Classification technique is the one which capable of processing a wider variety of data than regression and is growing in popularity.This can predict categorical class labels and that classifies data which based on training set and class labels. It is used for classifying the new data set. It is a part of data mining which gains more popularity.

Data mining involves the use of sophisticated data analysis tools which uses to discover the previously unknown, valid patterns and their relationships in large data set [1]. These tools can include the statistical models and the machine learning algorithms. The data mining consists of more than one collection and the managing data; it also includes analysis and the prediction.

There are several applications for Machine Learning algorithm and the significant techniques are definedin the mining. People are usuallymakes mistakes during the analyses or possibly when they trying to establish relationships between many features [3]. It is difficultto find the solutions for some problems. Machine learning can be successfully applied to these problems which help to improving the systems efficiency. There are many ML applications are involves in the tasks that can be set up as supervised.

In the present paper, I focused on the techniques which necessarily to do this. In particular, this work which is concerned with classification problems in where the output of instances admits only discrete and unlabeled values.

II. CLASSIFICATION TECHNIQUES

The operations of classification techniques have recently grown in advance.The popular methods as mentioned above were analyzed in detail.

A. K-Nearest Neighbor(K-Nn)

Nearest Neighbor (NN) is also known as ClosestPoint Search which used to identify the unknown data point that based on the nearest neighbor whose value is known. It has many applications in various fields such as Pattern recognition, Image databases and thebiomedical. The NN mechanism is classified into two different types such as Structure based and Structure less NN classification techniques. K-NN comes under the structure less classification technique. The structure based deals with the basic structure of the data where the structure has less mechanism which associated with training data samples [15]. Latter overcomes the memory limitation whereas the former reduces the computational complexity which makes use of the more than one nearest neighbor that determine the class in which the given data point belongs to and hence it is called as K-NN.

These data samples are needed to be in the memory at the run time where they are referred to as memory-based technique. All these data points are necessary that in order to make a decision which helps to determine the class of the given data point [36]. There are a large number of machine learning algorithms and K-NN is the most simplest among them. It can also be considered as the one among the top ten data mining algorithms.

K-NN basically works on the assumption that the datais contained in a feature space. Hence all the points aredefined in it, in order to find the distance amongthe points Euclidiandistance or Hamming distance isused according to the data type of data classes used.Here a single number “k” is given which is used todetermine the total number of neighbors that determinesthe classification. If the value of k is 1, it is also simplycalled as nearest neighbor. K-NN has following operations:

- An integer k
- A training data set
- A metric to measure closeness

The entiretechnique reviewed as determining the nearestneighbor and to finding its class using the neighborvalues.

B. *CART*

CART (Classification and Regression Tree)describes the high accuracy and handling noisy data or missing values. It takes arandom data which allows handling missing values by Chi-squared Automatic Interaction Detector(CHAID). In CART data pre-processing is not needed, it automatically selects relevant attributes [28]. CHAID algorithm considers missing values as distinctcategorical value, which also helps to method that are adopted. CART treats a refined method that ischanged, such as missing primaryfield. Itprunes to exact order that each nodemust be deleted. For small data sets once the Standard Error rule is good means it generates an optimal tree. For larger datasets zero Standard Error rulewhich generates the tree with high accuracy. Both C4.5 and CART are the robust tools. Surrogate lossfunction like Gini index is used when miss-classification of Decision tree.

C. Instance Based Learning

Instance based learning is the one which describes the lazy-learning algorithm, as a delay of induction or generalization process until aclassification is performed. Last learning algorithm is the term which requires less computation time during the training phase thaneager-learning algorithm (such as Decision tree & Bayesian network) that requires the more computation time during theclassification process [29].

It is called instance-based because it constructs hypotheses directly from the training data themselves. The computational complexity of classifying a single new data with instance by$O(n)$. The advantage of instance-based learning has over other methods of machine learning is its ability to adapt its model to previously unseen data: instance-based learners may simply store a new instance or throw an old instance away [32].The disadvantage of Instance Based Learning takes more computation time for data classification. Themodeling of input features through feature selection which improves classification accuracy and slow down classification time.

D. Support Vector Machines

Support Vector Machines is the one which uses to promising a new method for the Classification of both Linear and the Non Linear. This algorithm uses a Non Linear Mapping which uses to transform the original training data into a higher dimension [33]. This new dimension searches for Linear Optimal Separating Hyper plane (that is, a decision boundary separating the tuples of one class from another).The SVM is uses to finding thehyper planewhich describes Support Vectors (essential training tuples) and Margins [42].Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class also called functional margin and the larger margin with lower generalization errorof classifier.

E. Bayesian Network

Bayesian network is based on DAG (Directed Acyclic Graph) and one to one corresponding feature.Bayesian network is divided into two tasks learning DAG and design of network [6].The network design is fixed and the learning parameter in the conditional probability tables.Classifiers is the one that using Bayesian Classification which helps to predict the probability that a given tuple belongs to a single Class [8]. Baye's Theorem can predict the Posterior Probability, P(H,X) from P(H), P(X|H) and P(X).The X is a data tuple. Baye's Theorem is

P (H/X)=P(X/H) P(H) / P(X)

Where, H ->Hypothesis such as that the data tuple X belongs to a specified Class C
P(H|X) ->Probability that hypothesis H exists for some given values of X's attribute.
P(X|H) ->Probability of X conditioned on H
P(H) ->Probability of H
P(X) ->Probability of X

F. Minimum Description Length (Mdl)

Minimum description length(MDL) handles missing values naturally and chooses the missing values at randomly. The algorithm replaces sparse numericaldata with zeros and sparse categorical data with zero vectors [39]. Missing values arenested columns which are interpreted as sparse. The columns have missing data which are sample data types whichinterprets missing data at randomly.

MDL takes this into considerationof the size which model isreduction in uncertainty due to using the data model. Both entropy and model size are measured in data bits [40].The MDL mechanism is based on any regularity in a given set of data can be used to reduction of data than needed to describe the data literally.

III. EVALUATING THE PERFORMANCE OF CLASSIFIER

A. Hold- Out- method

The original data with labeled examples is classified into two different sets such astest set andtraining set. The data set would not be used in testing,the test set and learning [21]. Unseen test setprovides accuracy in unbiased estimation. It is mainly used for when the data set is large.

B. n-fold Cross-validation

The available data is partitioned into n equal-size disjoint subsets. Use each subset as the training set to learn a new classifier. The algorithm runn times, this gives n accuracy of the average accuracy. 10and 5-fold cross validations are the techniques which commonly used. This method is used when the data is not large.

C. Leave-one-out Cross validation

This method is used when the data set is very small. Hence it is a special case of cross-validation [26]. Each fold of cross validation which has only a single test example and all the test of the data are used in training. If the original data has m values, this is m-fold cross-validation.

D. Validation set

The available data is divided into three subsets,

1. Training set
2. Validation set and
3. Test set.

Avalidation set is used frequently for estimation parameters in Machine learning algorithm. The valuesthat give the best accuracy on the validation set are used as the final parameter values. Cross validation canbe used for parameter estimating as well.

IV. CONCLUSION

In this survey various techniques of Classification in Data MiningWere described in detail.These techniques are most important which uses to predict categorical class labels and that classify data that arebased on class labels and training set. It can be used for labeling andclassifying thenewly available data.Classification isused for the purposes of segmenting records. They have various achieve andobjectivesof their segmentations through various ways.

REFERENCE

[1] MichelineKamberand Jiawei Han,“Data Mining: Concepts and Techniques”,2nd Edition.

[2] S.Baik, J.Bala, “A Decision Tree Algorithm for Distributed Data Mining”,(2004).

[3] I.Witten & E.Frank, "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Francisco, 2nd Edition, (2005).

[4] Z.Zheng,“Constructing X-Of-N Attributes for Decision Tree Learning”. Machine Learning 40: 35–75, (2000).

[5] J.Sirgo, A. Lopez, R.Janez, R.Blanco, N.Abajo, M. Tarrio, R.Perez, “A Data Mining Engine Based OnInternet, Emerging Technologies And Factory Automation”.

[6] N.Friedman, D.Geiger&M.Goldszmidt,”Bayesian Network Classifiers”.Machine Learning 29:131-163, (1997).

[7] U.Fayyad, G.Piatetsky-Shapiro, And P.Smyth, “From Data Mining to Knowledge Discovery in Databases,” Ai Magazine,AmericanAssociation for Artificial Intelligence, (1996).

[8] N.Friedman&D.Koller,“Being Bayesian about Network Structure: A Bayesian Approach to StructureDiscovery in Bayesian Networks”. Machine Learning 50(1): 95-125, (2003).

[9] J.R.Quinlan, “C4.5 Programs for Machine Learning”.Morgan Kaufmann Publishers, San Francisco, Ca, (1993).

[10] D.L.Boley, ”Principal Direction Divisive Partitioning”. DataMining and Knowledge Discovery, 2, 4, 325-344, (1998).

[11] P.Bradley & U.Fayyad, “Refining Initial Points forK-Means Clustering”. In Proceedings of the 15th ICML, 91-99, Madison, WI, (1998).

[12] G.P.Babu & M.N.Marty, “Clustering with evolutionstrategies Pattern Recognition”, 27, 2, 321-329, (1994).

[13] T. Hastie, R. Tibshirani, J. Friedman, “The Elements of StatisticalLearning, Data Mining, Inference and Prediction”. Springer, New York,(2001).

[14] R.Duda&P.Hart, ”Pattern Classification and SceneAnalysis”. John Wiley & Sons, New York, NY, (1973).

[15] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda,G.J. McLachlan, A.F.M. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach,D.J. Hand, and D. Steinberg, “Top 10 Algorithms in Data Mining,”Knowledge and Information Systems, vol. 14, no. 1, pp. 1-37, (2008).

[16] Q. Yang and X. Wu, “10 Challenging Problems in Data MiningResearch,” Int’l J. Information Technology and Decision Making, vol. 5,no. 4, pp. 597-604, (2006).

[17] G.P.C. Fung, J.X. Yu, H. Lu, and P.S. Yu, “Text Classificationwithout Negative Examples Revisit,” IEEE Trans. Knowledge and DataEngineering, vol. 18, no. 1, pp. 6-20, (Jan 2006).

[18] H. Al Mubaid and S.A. Umair, “A New Text CategorizationTechnique Using Distributional Clustering and Learning Logic,”IEEE Trans. Knowledge and Data Eng., vol. 18, no. 9, pp. 1156-1165, (Sept 2006).

[19] Gregory. Piatetsky, “Data mining and knowledge discovery 1996 to 2005: overcoming the hype and moving from university to business and analytics. Data Mining Knowledge Discovery (2007)15:99–105,(2007).

[20] Hans-Peter.Kriegel, Karsten M.Borgwardt &Peer. Kröger, “Future trends in data mining”. Data Mining Knowledge Discovery 15:87–97, (2007).

[21] Qi.Luo, “Advancing Knowledge Discoveryand Data Mining” Knowledge Discovery and Data Mining, WKDD (2008).

[22] Gary M.Weiss, Bianca.Zadrozny, Maytal.Saar-Tsechansky, “Guest editorial: special issue onutility-based data mining”. Data Mining Knowledge Discovery 17:129–135, (2008).

[23] Ben G.Weber &Michael.Mateas, “A Data Mining Approach to Strategy Prediction” 978-1- 4244-4815, IEEE, (2009).

[24] Sufal.Das &Banani.Saha,“Data Quality Mining using Genetic Algorithm”. International Journal of Computer Science and Security, (IJCSS) Volume (3): Issue (2), (2009).

[25] Atul. Kamble,"Incremental Clustering in Data Mining using Genetic Algorithm". International Journal of Computer Theory and Engineering, Vol.2, No. 3, (June 2010).
[26] Murat.Kantarcıoglu, Bowei.Xi &Chris.Clifton, "Classifier evaluation and attribute selection against active adversaries" Data Mining Knowledge Discovery 22:291–335, (2011).
[27] R.Bouckaert, "Naive Bayes Classifiers That Perform Well with Continuous Variables, Lecture Notes in Computer Science". Volume 3339, Pages 1089 – 1094,(2004).
[28] L. A.Breslow,&D.W.Aha, "Simplifying decision trees: A survey". Knowledge Engineering Review 12: 1–40, (1997).
[29] H.Brighton &C.Mellish, "Advances in Instance Selection for Instance-Based Learning Algorithms". DataMining and Knowledge Discovery 6: 153–172, (2002).
[30] Y.Yang & G.Webb,"On Why Discretization Works forNaive-Bayes Classifiers, Lecture Notes in Computer Science", Volume 2903, Pages 440 – 452,(2003).
[31] G.Zhang, "Neural networks for classification: a survey". IEEE Transactions on Systems, Man, and Cybernetics, Part C30(4): 451- 462, (2000).
[32] D. R.Wilson & T.Martinez, "Reduction Techniques for Instance-Based Learning Algorithms". Machine Learning 38:257–286,(2000).
[33] Danny Roobaert. "DirectSVM: A fast and simple support vector machineperceptron". In Proceedings of IEEE International Workshop on NeuralNetworks for Signal Processing, Sydney, Australia, (December 2000).
[34] M. Narasimha Murty andS. V. N. Vishwanathan,"Geometric SVM: A fast and intuitive SVM algorithm". Technical Report IISC-CSA-2001-14, Dept. of CSA, Indian Institute of Science, Bangalore, India, November 2001. Submitted to ICPR (2002).
[35] J. P. Zhang, L. L. Chen, J. Yang, and J. Ma, "ICA-basedattribute Bagging support vector machine integration method", Journal of Dalian Maritime University, Vol.34, No.3, 125-127, (2008).
[36] Dr.K.Thanushkodi andN.Suguna, "An Improved k- Nearest Neighbour Classification Using Genetic Algorithm", International Journal of Computer Science Issues, vol. 7, no.2, pp. 18-21, july (2010).
[37] M. Kubat, "Voting Nearest NeighbourSubclassifiers". Proceedings Of The 17thInternational Conference Onmachine Learning, ICML-2000, Pp.503-510, Stanford, CA, June 29-July 2, (2000).
[38] R. C.Barros, R.Cerri, P. A.Jaskowiak, A. C. P. L. F.Carvalho, "A Bottom-Up Oblique Decision Tree Induction Algorithm". Proceedings of the 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011).
[39] P. Grünwald,"Advances in Minimum Description Length: Theory and Applications". MIT Press. Retrieved 2010-07-03, (April 2005).
[40] S.Argamon, "Efficient Unsupervised Recursive Word Segmentation Using Minimum Description Length". Proc. 20th International Conference on Computational …. (2.26/year),et al., (2004).
[41] Finn V. Jensen, Thomas D.Nielsen, "Bayesian Networks and Decision Graphs". Information Science and Statistics series (2nd ed.). New York: Springer-Verlag. ISBN 978-0-387-68281-5,(June 6, 2007).
[42] H. William,Saul A.Teukolsky, William T.Vetterling and B. P.Flannery,"Section 16.5. Support Vector Machines". Numerical Recipes: The Art of Scientific Computing (3rd ed.). New York: Cambridge University Press. ISBN 978-0-521-88068-8, (2007).
[43] Y.Lee, Y.Lin, G.Wahba, "Multicategory Support Vector Machines". Journal of the American Statistical Association99 (465): 67. doi:10.1198/016214504000000098, (2004).