IJIRMET
ISSN (Online): 2456-0448
International Journal Of Innovative Research
In Management, Engineering And Technology
Available online

A REVIEW OF DATA MINING TECHNIQUES CLASSIFICATION AND FEATURE SELECTION

[1] *Jency.W.G*
Research Scholar, Noorul Islam Centre for Higher Education,Thuckalay, Kumaracoil, Tamil Nadu 629180
Asst. Professor,Department Of Computer Science and Engineering,
John Cox Memorial CSI Institute of Technology-Kannammoola, Thiruvananthapuram,Kerala 695011
jencywg@gmail.com

[2] *Dr.J.E.Judith*
Associate Professor, Department Of Computer Science and Engineering,Noorul Islam Centre for Higher Education, Thuckalay, Kumaracoil, Tamil Nadu 629180
judithjegan@gmail.com

To access & cite this article
Website: www.ijirmet.com

ABSTRACT

As the world grows in complexity and huge amount of data generates time to time, data analysis become difficult. In recent years, large amount of data are collected for research purposes. Such data set consists of hundreds or thousands of features. Many of the features in such data are useful information relevant to the problem. It also contains irrelevant information. So to extract relevant information a pre processing step called Feature Selection is used. Feature selection techniques like wrapper, filter, and embedded techniques are used. In Feature Selection process the relevant data are filtered to reduce the complexity before applying data mining techniques. Data mining is the process of discovering hidden, previously unknown and useful patterns essential for solving problems. For discovering classes of unknown a data mining technique called Classification is used. There are different method for classification like Bayesian, decision trees, rule based, neural networks etc. This paper analysis some existing and popular feature selection algorithms and classification.

KEYWORDS: Data Mining, Feature Selection, Classification

Copyright to IJIRMET

I. INTRODUCTION :

With the growing demand in the day today life the data mining has become popular in recent years. Data mining is knowledge discovery from database. Data mining is the process of extracting knowledge from huge amount of data [1]. Data mining consists of several steps. The process of extracting knowledge consist of the steps namely data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation and knowledge discovery. The hidden data is collected in the form of patterns, concept, rules and so on [2].

A Pre-processing step called feature selection is used to reduce the dimensionality before applying any data mining techniques. For Feature selection, techniques like wrapper, filter, embedded or hybrid method is used. According to different application the mining task can be mainly divided into class or concept description, association analysis, classification or prediction and clustering. Various clustering method existing is Bayesian, decision trees, rule based, neural networks etc

The task in data mining can be classified into two: descriptive and predictive. Descriptive mining characterize the general properties of data in the database. Predictive mining inference is made on the current data to make predictions.

This paper provides a survey on various feature selection techniques and classification techniques used in data mining.

II. FEATURE SELECTION METHODS:

Feature selection algorithms are broadly classified into three categories namely Filter, Wrapper and Hybrid method [3].

Filter Method selects the feature subset on the basis of intrinsic characteristics of the data, independent of mining algorithm. It can be applied to data with high dimensionality. The advantages of Filter method are its generality and high computation efficiency.

Wrapper Method requires a predetermined algorithm to determine the best feature subset. Predictive accuracy of the algorithm is used for evaluation. This method guarantees better results, but it is computationally expensive for large dataset. For this reason, the Wrapper method is not usually preferred [4].

Hybrid Method combines Filter and Wrapper to achieve the advantages of both the methods. It uses an independent measure and a mining algorithm to measure the goodness of newly generated subset [5]. In this approach, Filter method is first applied to reduce the search space and then a wrapper model is applied to obtain the best feature subset [6].

IV. COMPARISON OF FEATURE SELECTION ALGORITHMS :

Feature selection is the essential preprocessing step in Data mining. Several feature selection algorithms are available. Each algorithm has its own strength and weakness. Table 1 compares some of the available algorithms.

Algorithm	Type	Benefit	Drawback
Relief [7]	Filter	It is scalable to data set with increasing dimensionality.	It cannot eliminate the redundant features.
Correlation-based Feature Selection [8]	Filter	It handles both irrelevant and redundant features and It prevents the reintroduction of redundant features.	It works well on smaller datasets It cannot handle numeric class problems.
Fast Correlation Based Filter [3]	Filter	It hugely reduce the dimensionality	It cannot handle feature redundancy.

Copyright to IJIRMET

Interact [9]	Filter	It improves the accuracy.	Its mining performance decreases, as the dimension-ality increases.
Fast Cluster-ing-Based Feature Sub-set Selection [6]	Filter	Dimensional-ity is hugely re-duced	Works well only for Microar-ray data.
Condition Dynamic Mutual In-formation Feature Se-lection [10]	Filter	Better Performance	Sensitive to noise
Affinity Propagation – Sequential Feature Selection [11]	Wrap-per	Faster than Sequential Feature Selection	Accuracy is not better than SFS
Evolution-ary Local Selection Algorithm [12]	Wrap-per	Covers a large space of possible feature combinations	As the num-ber of features increases, the cluster quality decreases.
Wrapper Based Fea-ture Selection using SVM [13]	Wrap-per	Better Accu-racy and Faster Computation	
Two-Phase Feature Selection Approach [14]	Hyb rid	Handles both irrelevant and Redundant features. Improves Accuracy	
Hybrid Feature Selection [15]	Hyb rid	Improves Accuracy	High Com-putation Cost for high dimensional data set

Table 1: Comparison of Existing feature selection algorithms

Filter methods are much faster and better than wrappers. It can be applied to large datasets having many features [16]. But Filter Method is not always enough to obtain better accuracy [17]. On the other hand, Wrapper Method also selects best feature subsets but it has proven to have high computation cost when compared to Filter for large datasets [16]. Hybrid method is less computationally intensive than wrapper methods.

V. CLASSIFICATION :

Classification is a technique used for discovering unknown data. Table 2 shows the various methods of classification and approaches used.

Classifications	Approaches used
Rule Based Classifi-er[18]	If-then rules
Bayesian Net-work[19]	Directed ,acyclic graph and probabili-ty distribution
Decision Tree[20]	Root-test, leaf-class-es for the instance
Nearest Neigh-bor[21]	Greater weight are given to closer points
Artificial Neural Net-works[22]	Input layer, Hidden layer, weights, out-put layer
Support Vector Ma-chine[23]	Statistical learning theory, probability
Rough sets[24]	Lower and upper approximation
Fuzzy Logic[25]	Fuzzy logic vari-ables, range between 0 and 1
Genetic Algo-rithm[26]	Natural genetics search processes

Table 2: Various methods of classification and approaches used.

IV. REFERENCES :

1. J. Han and M. Kamber, *Data mining concepts and techniques,* Morgan Kaufmann, San Francisco 2006

Copyright to IJIRMET

2. T.J. Shan, H. Wei and Q. Yan, "Application of genetic algorithm in data mining", *1st Int Work Educ Technol Comput Sci, IEEE 2*,2009,pp.353-356
3. Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Department of Computer Science & Engineering, Arizone State University, Tempe, AZ 85287-5406, USA, 2003
4. A.Blum and P.Langley, "Selection of relevant features and examples in machine learning", Artificial Intelligence, vol 97, pp 245-271,1997
5. Huan Liu and Lei Yu, " Towards Integrating Feature Selection Algorithms for Classification and Clustering", IEEE Transactions on Knowledge and Data Engineering, Vol.17 No.4 2005
6. Qinbao Song, Jingjie Ni and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, Vol 25, No.1, 2013
7. K.Kira and L.A Rendell, "The Feature Selection Problem: Traditional methods and A New Algorithm," Proc. 10th National Conference Artificial Intelligence, pp.129-134, 1992.
8. Mark A. Hall and Lloyd A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper", Proceedings of the Twelfth International FLAIRS Conference, 1999.
9. Zheng Zhao and Huan Liu "Searching for Interacting Features" Department of Computer Science and Engineering, Arizona State University, 2007
10. Wang Liping, "Feature Selection Algorithm Based On Conditional Dynamic Mutual Information", International Journal O Smart Sensing and Intelligent Systems", VOL. 8, NO. 1, 2015
11. Kexin Zhu and Jian Yang, "A Cluster-Based Sequential Feature Selection Algorihm", IEEE, 2013
12. Y.Kim, W.Street, and F.Menczer, "Feature Selection for Unsupervised Learning Via Evolutionary Search," Proc. Sixth ACM SIGKDD International Conference, Knowledge Discovery and Data Mining, pp 365 – 369, 2000
13. Hwang, Young-Sup, "Wrapper-based Feature Selection Using Support Vector Machine", Department of Computer Science and Engineering, Sun Moon University, Asan, Sunmoonro 221-70, Korea, Life Science Journal 2014;11 (7)
14. B.M Vidhyavathi, " A New Approach to Feature Selection for Data Mining", International Journal of Computational Intelligence Research, ISSN 0973-1873 Vol.7 Number 3, pp 263 – 269, 2011
15. Jihong Liu, "A Hybrid Feature Selection Algorithm for Data sets of thousands of Variables" IEEE, 2010
16. Mark A. Hall and Lloyd A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-based Filter Approach to the Wrapper", Proceedings of the Twelfth International FLAIRS Conference, 1999.
17. Dr. S. Rabiyathul Basariya, and Dr. Ramyar Rzgar Ahmed, 2019. "The Influence of 'Adventure Tourism Activities' in promoting tourism business in mountain stations", African Journal of Hospitality, Tourism and Leisure, Volume 8 (2).
18. Dr. S. Rabiyathul Basariya, and Dr. Ramyar Rzgar Ahmed, Nov 2018. "A Study On consumer satisfaction and preference of colour TV brands in Chennai city", International Research Journal of Management and Commerce, Volume4, Issue 10.
19. Dr. S. Rabiyathul Basariya, and Dr. Ramyar Rzgar Ahmed, "A Study on Attrition: Turnover intentions of employees", Jan 2019. International Journal of Civil Engineering and Technology (IJCIET), Volume 10, Issue 9.
20. Dr. S. Rabiyathul Basariya, and Dr. Nabaz Nawzad Abdullah, Dec 2018. "A STUDY ON CUSTOMER'S SATISFACTION TOWARDS E-BANKING", International Research Journal of Management and Commerce, Volume 5, Issue 12,
21. Thomas Joy, Basariya S. Rabiyathul, "A study on the issues of Financial ratio analysis" Indian Journal of Public Health Research & Development, 2019, Volume : 10, Issue : 3, pp 1079-1081
22. B.M Vidhyavathi, " A New Approach to Feature Selection for Data Mining", International Journal of Computational Intelligence Research, ISSN 0973-1873 Vol.7 Number 3, pp 263 – 269, 2011
23. Thomas Weise, Raymond Chiong, "Evolutionary data mining approaches for rule based and tree-based

Copyright to IJIRMET

classifiers",9th IEEE International Conference on Cognitive Informatics(ICCI'10),2010,pp. 696 - 703

24. G.F.Cooper,P.Hennings-Yeomans,S.visweswaran and M.Barmada, "An Efficient Bayesian Method for Predicting Clinical Outcomes from Genome-Wide Data", AMIA 2010 Symposium Proceedings,2010, pp 127-131
25. M.Garofalakis, D.Hyun, R.Rastogi and K.Shim "Building Decision Trees with Constraints", Data Mining and Knowledge Discovery, vol 7,no 2,2003,pp187-214
26. T.M.Mitchell, Machine learning, McGraw Hill Companies, USA, 1997
27. Y.Singh, A.S.Chauhan, "Neural Networks in Data Mining", Journals of Theoretical and Applied Information Technology, 2005, pp.37-42
28. V.N.Vapnik, Statistical Learning Theory,Wiley New York,1998
29. Z.Pawlak,"Rough sets", International Journal of Computer and Information Sciences,1982,pp.341-356
30. L.Tari,Cbaral and S.Kim, "Fuzzy c-means clustering with prior biological knowledge", Journal of Biomedical Informatics, 42(1),2009,pp.74-81
31. D.E.Goldberg, Genetic algorithms in search optimization and machine learning, Newyork, 1989

Copyright to IJIRMET