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 8, August 2016

Improving Medical Diagnosis using Filter and Wrapper Techniques

Sonu Rani[1], Dharminder Kumar[2], Sunita Beniwal[3]

M. Tech Scholar, Computer Science &Engg, Guru Jambheshwar University of Science & Technology, Hisar, India[1]

Professor, Computer Science &Engineering, Guru Jambheshwar University of Science & Technology, Hisar, India[2]

Assistant Professor, Computer Science &Engg, Guru Jambheshwar University of Science & Technology, Hisar, India[3]

Abstract: Data mining also called knowledge discovery is essential for solving problems in specific domain. Medical diagnosis is a time consuming and difficult task performed by experts. For improving medical diagnosis using data mining techniques, data mining techniques need to be applied on relevant features. Over fitting can be reduced and performance of mining can be enhanced if more relevant features are selected. This paper reviews the various feature selection techniques and their application in different domains.

Keywords: Relief Attribute Evaluation, Symmetric uncertainty attribute, Multilayer Perceptron

I. INTRODUCTION

Data Mining is a young and promising field in the information industry. It is a natural evolution of database system technology. The data is increasing day by day so the situation is "Data Rich Information Poor"[1]. Researchers are continuously developing tools and technology to turn data into information.

In large datasets the problem is how to analyze large amounts of data. Here the data mining concepts and techniques helps to uncover interesting hidden data patterns from huge amount of data.
Data Mining refers to extraction of novel, interesting and valid information from the huge data. In the process of knowledge discovery data mining is the most essential step. Data Mining helps analyst to solve problems in specific domain. It also helps in the process of decision making[2].

II. FEATURE SELECTION

Feature Selection is a process in which the features that are relevant to the application domain are retrieved by applying feature selection techniques. The large dataset contains raw data with many irrelevant attributes. The irrelevant attributes may degrade the performance of data mining tasks and techniques such as classification, clustering etc.

So, irrelevant attribute needs to be filtered to increase the efficiency and accuracy of such tasks[3]. Feature selection is an amazing pre-processing technique that can do the task accurately by giving the subset of features that are relevant to specific domain. The aim of attribute selections to enhance the model performance to provide fast and cost effective models for mining. Broadly feature selection techniques can be categorized into three categories[4]

A. Filter Approach
B. Wrapper Approach
C. Embedded Approach

The Selection of relevant optimal subset of features may add complexity to the model, so the approach used should be cost effective also. Pre-processing improve the efficiency and ease of mining process and also decrease the computational cost. Attribute subset selection reduces the size of data sets by removing redundant and irrelevant attributes from the data set[5].

The best or worst attributes/features are selected by performing statistical test to determine the significance of attributes in the specific domain. Many other attribute evaluation measures can be used such as information gain, gain ratio, PCA, Relief Attribute Evaluator etc.[6]

A. The Filter Approach

In Filter selection approach, features are selected before the induction step. The filter approach for feature selection is independent of the mining algorithm that is applied to all the attributes in dataset. The relevance of features in this method is determined by intrinsic properties of data.[9]The features subset selected in this way are given as input to the classification algorithm.

The advantage of feature selection approach is that they are easily scalable to high dimensional datasets, simple, cost effective and feature selection is performed only once and then different classifiercan be used for evaluation.

Disadvantages of filter approach are that here is no interaction with the classifier that is used for evaluation, thus there is no feature dependencies due to which the performance of classification algorithm get affected.[10]

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 8, August 2016

B. The Wrapper Approach

The wrapper approach of feature selection makes use of data mining algorithm to check the goodness of the subset selected. A search method is used for possible subset of features.[11] In this approach feature subset are generated and evaluated. When comparing to filterapproach the wrapper mining algorithms is applied to eachwhen comparing to filter approach the wrapper mining algorithms is applied to each attribute subset generated by the search method. Advantages of wrapper approaches are that there is interaction between feature subset search and the model selected for classifier. Disadvantage of wrapper approachis that it is computationally expensive and also higher risk of over fitting[12]

C. Embedded Approach

In this approach the filter technique is incorporated into the classifier itself. As the approach for selection of optimal subset of features is into the classifier itself, theapproach is specific to the data mining learning algorithm. The advantages of both filter and wrapper approach are combined in this interaction withthe classification model and also less computationally intensive.[13]

III. LITERATURE REVIEW

Lots of work has been done on feature selection techniques for selecting optimal attribute subset. Medical datasets available comprise of many attributes. Many of attributes may not be useful for decision making. Many techniques have been proposed for feature selection.

Unler et al., [14] proposed a hybrid model for feature selection which incorporates both filter and wrapper techniques. On medical data sets firstly they applied feature selection techniques to reduce irrelevant attributes and then they apply wrapper approach to reduce the cost theatrically.

Zhang et al.,[15] used rough set theory for relevant feature selection and proposed an incremental method for vital data mining. Through rough sets they defined amalgamated information systems that contained attributes of collaborative different types, whichwas subject to for feature selection and knowledge discovery. Li et al., 2010, [16]they proposed a distributed and parallel Genetic Algorithm for feature selection.

Tsai and Hsiao et al., 2010, [17] proposed a filter technique by assimilate PCA(Principal Component Analysis),Decision trees(CART) and Genetic algorithms(GA).The proposedmethod filtered out extraneous variables on union, intersection, and multi-intersection strategies.

Uguz et al., 2011,[18] devise the feature selection avenue for text categorization and performed feature selection in two stages First ranker algorithm is applied to assign rank to each term in the document depending on their significance for classification using entropy. In the next stage GA (Genetic Algorithm) and (PCA) Principal Component Analysis are applied respectively to the terms ranked into the document.

Qaunz et al., 2012,[19] propound a unprecedented feature selection technique for real and synthetic data using sparse coding approach.

Pacheco et al., 2013,[20] proposed a unprecedented method NSGAFS(non-dominated sorting genetic algorithm)for feature selection.

Sun et al., 2013,[21] they proposed a dynamic weighting-based feature selection algorithm thatassign ranks tofeatures based on information metric.

Chen et al., 2009,[22]proposed a Semantic Relationship Graph (SRG) to describe the relation between multiple tables and the search for pertinent features performed within the relational space.

Peker et al .,2015[23] proposed effective feature selection using algorithms such as minimum redundancy maximum relevance , ReliefF, and Sequential Forward Selection.

Noruzi et al., 2015[24] proposed a graph based feature selection for improving medical diagnosis using graph clustering.

Ramchandra et al., 2014[25] proposed feature selection method for large medical data sets using SVM with genetic algorithm.

IV. CONCLUSION

From the literature review we concluded that a lot of work has been done on medical data set. But there is a lot to explore there are so many techniques that can be applied. The basic technology that is applied on data sets is feature selection because feature selection can minimize the task a lot but there are so many feature selection techniques each have some pros and cons .So from the literature review we got the idea to combine the pros of techniques and make a technique that is combination of filter and wrapper approaches of feature selection and then apply combination of filter and wrapper approach on medical data sets and then compare their efficiency.

REFERENCES

[1] J. Han, M. Kamber, and J. Pei, “1 - Introduction,” in Data Mining (Third Edition), Boston: Morgan Kaufmann, 2012, pp. 1–38.
[2] J. Han, M. Kamber, and J. Pei, “2 - Getting to Know Your Data,” in Data Mining (Third Edition), Boston: Morgan Kaufmann, 2012, pp. 39-82
[3] S. Beniwal and J. Arora, “Classification and feature selection techniques in data mining,” In International Journal of Engineering Research & Technology, vol. 1, no. 6, 2012, pp. 1-6.
[4] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining. Springer Science & Business Media, 2012.
[5] Y. Saeys, I. Inza, and P. Larrañaga, “A review of feature selection techniques in bioinformatics,” Bioinformatics, vol. 23, no. 19, 2007, pp. 2507–2517.
[6] H. Liu and H. Motoda, Computational Methods of Feature Selection. CRC Press, 2007.

[7] L. C. Molina, L. Belanche, and A. Nebot, "Feature selection algorithms: a survey and experimental evaluation," in IEEE International Conference on Data Mining, 2002, pp. 306–313.
[8] A. G. Karegowda, M. A. Jayaram, and A. S. Manjunath, "Feature subset selection problem using wrapper approach in supervised learning," International journal of Computer applications, vol. 1, no. 7, 2010, pp. 13–17.
[9] R. Wald, T. Khoshgoftaar, and A. Napolitano, "Comparison of Stability for Different Families of Filter-Based and Wrapper-Based Feature Selection," in Machine Learning and Applications, 12th International Conference on, vol. 2, 2013, pp. 457–464.
[10] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on, 2012, pp. 356–363.
[11] A. Abu Shanab, T. M. Khoshgoftaar, and R. Wald, "Evaluation of Wrapper-Based Feature Selection Using Hard, Moderate, and Easy Bioinformatics Data," in Bioinformatics and Bioengineering (BIBE), 2014 IEEE International Conference on, 2014, pp. 149–155.
[12] Y. Han, K. Park, and Y.-K. Lee, "Confident wrapper-type semi-supervised feature selection using an ensemble classifier," in Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on, 2011, pp. 4581–4586.
[13] G. Li, X. Hu, X. Shen, X. Chen, and Z. Li, "A novel unsupervised feature selection method for the data sets of bioinformatics through feature clustering," in Granular Computing, 2008. GrC 2008. IEEE International Conference on, 2008, pp. 41–47.
[14] A. Unler, A. Murat, and R. B. Chinnam, "mr 2 PSO: a maximum relevance minimum redundancy approach based on swarm intelligence for support vector machine classification," Inf. Sci., vol. 181, no. 20, pp. 4625–4641, 2011.
[15] J. Zhang, T. Li, and H. Chen, "Composite rough sets for dynamic data mining," Inf. Sci., vol. 257, pp. 81–100, 2014.
[16] R. Li, J. Lu, Y. Zhang, and T. Zhao, "Dynamic Adaboost learning with feature selection that is based on technique of parallel genetic algorithm for image annotation," Knowledge.-Based System., volume. 23, no. 3, pp. 195–201, 2010.
[17] C.-F. Tsai and Y.-C. Hsiao, "Combining multiple feature selection methods for stock prediction: By using Union, intersection, and multi-intersection approaches," Decision. Support System., volume. 50, no. 1, pp. 258–269, 2010.
[18] H. Uğuz, "A two-stage feature selection method used for text categorization by using information gain, principal component analysis(PCA) and genetic algorithm(GA)," Knowledge.-Based System., volume. 24, no. 7, pp. 1024–1032, 2011.
[19] B. Quanz, J. Huan, and M. Mishra, "Knowledge transfer with low-quality data: A feature extraction issue," Knowledge of Data Engineering & IEEE Transaction. On, vol. 24, no. 10, pp. 1789–1802, 2012.
[20] S. Casado, F. Angel-Bello, A. Álvarez, and others, "Bi-objective feature selection for discriminant analysis in two-class classification," Knowledge.-Based System., volume. 44, pp. 57–64, 2013.
[21] J. Sun and A. Zhou, "Unsupervised robust Bayesian feature selection," in Neural Networks (IJCNN), 2014 International Joint Conference on, 2014, pp. 558–564.
[22] H. Chen, H. Liu, J. Han, X. Yin, and J. He, "Exploring optimization of semantic relationship graph for multi-relational Bayesian classification," Decision. Support System., volume. 48, no. 1, pp. 112–121, 2009.
[23] M. Peker, A. Arslan, B. Sen, F. V. Çelebi, and A. But, "An unprecedented hybrid method for determining the depth of anesthesia level: Combining ReliefF feature selection approach with random forest algorithm (ReliefF+RF)," in Innovations in Intelligent SysTems and Applications (INISTA), 2015 International Symposium on, 2015, pp. 1–8.