Analysis of Feature Selection Algorithms and a Comparative study on Heterogeneous Classifier for High Dimensional Data survey

Kassahun Azezew Ayidagn[1], prof. Shilpa Gite[2]
Department of Computer science/Information Technology
Symbiosis Institute of Technology, Pune, India

Abstract: *This paper focuses on the analysis of various feature selection algorithms and a comparative study on heterogeneous classifier predictive accuracy problems to work with high dimensional data. Especially we conduct experimental comparisons of IBK (KNN), SVM, NBTree and J48 on KDD Cup99 intrusion detection dataset and one cancer disease diagnosis microarray datasets and analysis their performance with vote generalizations. Based on the fact a large number of features can cause a noise of data and degrades a performance of learning algorithm.To tackle these problems identifying a suitable feature selection method is essential for a given machine learning algorithm tasks. So feature selection plays a great role in intrusion detection, bioinformatics, and medical data analysis. Thus this paper deals the application of best feature selection techniques to improve learning algorithm predictive accuracy in microarray dataset and KDD (Knowledge Discovery and Data Mining Tools Conference) Cup 99 dataset with a respective classification and feature selection algorithms. basically, this approach shows the application of feature selection algorithms when a large number of features represented in a small sample data and small numbers of features represented with a high number of samples by taking the above two different datasets.*

Keywords: *High dimensional data. feature selection algorithm. Heterogeneous classifier. Feature selection.*

1. Introduction

In advanced technology handling a large amount of data is a challenging task among researchers since the data are stored through various data acquisition techniques. This accumulated massive amount of data decreases the learning algorithm performance in terms of causing overfitting, increasing learning module complexity since raw data have a number of a feature known as high dimensional data. In the fact, high dimensional data contains irrelevant data and redundant features which increases the noise of data and error of learning algorithms. Feature selection is a solution for a such like problems.it reduces data dimensionality by using different feature selection techniques. there are three main feature selection categories namely, filter, wrapper and embedded methods. Many concepts of these methods are found in the literature. Asir et al. (2016) is a good example of reviews. The process of feature selection categorizes into two namely feature ranking method and feature subset selections based on how the feature combined with different feature selection algorithms for evaluation[15]. Filter method namely Chi-square, info-gain, MrMr (minimum redundancy maximum relevance) selects the feature based on feature relevance and embedded methods namely SVM-RFE (recursive feature elimination for support vector machine). Selects best feature subsets based on usefulness criterion. Thus feature evaluator ranks the results based on feature relevance or usefulness and those results are combined into one before training on classifiers in order to increase learning algorithm performance and feature selection stabilities. As the feature evaluator ensemble increases the test consistency and learning algorithms ensemble also increase the model generalizability and decreases model complexity. The ensemble can be done in the feature selector level and classification levels. The ensemble can be heterogeneous and homogeneous. Heterogeneous is a different type of base learners trained on the same sample data and homogenous means the same type of base learners trained on different sample data [4].

The purpose of this introductory section is to briefly describe the challenges of feature selection in high dimensional data and to suggest the latest technique that helps to address those problems. Based on the fact when the number of features or samples increased the learning algorithm face challenges to cop up with those data. So it needs to pay more attention in feature selection algorithms and classification algorithms. The dimensionality is measured by a number of features or attributes and the number of samples or instances. So the proposed approach investigates a high number of features represented by a low number of sample in microarray datasets and a low number of attributes or features represented in a high number of samples in KDD Cup99 intrusion detection datasets. To work on the above two datasets we use embedded and filter feature selection methods and ensemble techniques namely vote. A vote is an ensemble technique that

helps to combine different classifiers into one and train on a given datasets. Models were combined using a simple voting mechanism, with each algorithm model having one vote. To break ties, however, a slight weighting factor was used: model weights were created so that the models that performed best during training were given slightly more weight than others. The four algorithms used were SVM (support vector machine), J48, k-nearest neighbor, NBTree. And we conduct experimental comparisons on those learning algorithms and analyze their performance with vote generalization.

Basically, in this study, we are using the data mining approaches classification, and attribute selection with respective data pre-processing namely data cleaning, data transformation, data discretization, and data integration. Since we are dealing about high dimensional data discretization is more necessary to work in KDD Cup99 intrusion detection system datasets. And data cleaning also applying in the process filling in the missing value, smoothing the noisy data or resolving the inconsistencies in the data. Basically, this work aims to provide a deep study on the advantages and disadvantages of different existing ensemble classifiers and feature selection algorithms. In this study, we use a Data Mining tool Weka for classification and attribute selection and classifier ensemble methods namely vote.

2. Related work

In the past decades, many researchers have devoted ensemble learning algorithms in order to work with high dimensional data.

The problems of feature selection algorithms to work with high dimensional data comprehensively defined by Veronica et al. (2016). Those are dataset shift, incremental learning, and class imbalance. The study basically defines the categories of the feature selection methods from the functional point of views namely ranker and subset and from a structural point of views embedded filter and wrapper methods. In The functional point of view, the feature selection works in two different ways [3]. Some of the feature selection methods are assigning weights to each feature with its relevance measure. On the other hand, the subset features are generated based on its usefulness measure. The classification algorithms like SVM and C4.5 in microarray datasets and SVM show a higher performance over C4.5 [3]. On the other hand, C4.5 show a better performance in KDD Cup99 intrusion detection datasets compared with the existed methods [3]. The KDD Cup dataset contains five million samples represented by 41 features with the aim of categorizing each connection in one of the following class: denial of service (DoS), Probe attack, Remote-to-Local (R2L) attacks and User-to-Root (U2R) attacks [3].

Heterogeneous and homogenous ensembles methods are proposed by Afefben et al. (2017). These methods are new methods that help to keep the stability of feature selection and classifier predictive accuracy. The same types of base learners train on different sample data is a heterogeneous ensemble and different types of base learners train on the same sample data is heterogeneous ensemble [4]. Then they propose robust aggregation based on classification performance and reliability assessment to combine selector's ensemble output. To ensemble the resulting feature list they proposed different aggregation techniques [4]. These techniques are weighted mean aggregation, complete linear aggregation, robust rank aggregate, feature occurrence frequency and classification accuracy based aggregation. An ensemble feature selection with reliability assessment and robust aggregation always improves the predictive accuracy of learning algorithms [4]. The experiments showed as that robust ensemble aggregation method improves the classification performance and stabilities of feature selection for seven high dimensional cancer diagnosis microarray datasets [4]. They measured the stability on the different selected feature subsets find from the tenfold cross-validation and they used the KNN classifier to build a classification model. Thus the KNN classifier predictive accuracy on the proposed approach shows the improved results.

So far researchers proposed a number of different feature selection methods namely filter based, wrapper based and embedded based [15], [10], [19]. Filter methods build the selected feature subsets independently of the learning algorithms by using the training data**.** In Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. Wrapper method uses the learning algorithms to score feature subsets [19]. Embedded methods learn which features best contribute to the accuracy of the model while the model is being created.

Perthame et al. (2015). Investigates the heterogeneity of data affects the ranking and stability of supervised classification model selection. In a high dimensional data feature selection is a key point to address dimensionality problems [3], [4], [8]. The dimension increases the learning algorithm performance decreases. To tackle this problem a number of feature selection and classification techniques are proposed [4], [16]. Those are ensemble or hybrid methods. These methods are applied on the classification level and feature selection levels with the aim of improving the stability of feature selection and classifier predictive accuracy.

Stability of feature selections methods and different stability measures are defined in David et al. **(2014).** The stability of feature selection is keeping the consistency of the selector test performance. Different stability measures that are applied in this

papers are: feature focussed versus subset focussed measures, selection register versus selection exclusion register measure and subset biased versus subset –size-unbiased measures.

The studies of the NSL-KDD dataset in a class-wise analysis are proposed by Aggarwal et al. (2015). That class is Basic, Content, Host, and Traffic. Basic means features are the feature of each TCP connection, Content means features are the attributes within a connection provided by the domain knowledge, Traffic means features are the attributes computed using a two-second time window, Host means features are the attributes design to assess attacks which last for more than two seconds. This class wise analysis is proposed with the aim of representing 42 attributes and categorizing the connection in the above mentioned four classes. Intrusion detection metric helps to measure an intrusion detection system [1]. Some of the evaluation metrics are false alarm rate, detection rate, accuracy, precision, specificity, and F-score. These evaluation metrics are derived from the four basic attributes the confusion matrices. Those attributes are true negative, false negative, false positive and true positive. In the experimental result analysis the content class attributes are present in combination with other class attributes are the detection rate is above 70% almost all the combination [1].in the KDD Cup99 Intrusion detection datasets, they used to build classification models and the performance is evaluated in terms of AUC performance metric. Thus the result analysis showed as the detection rate increased and the false alarm rate decreased.

Different classification algorithms are defined by Nagi et al. (2013). These classification algorithms are Naïve Bayes: it's a probabilistic classifier based on Bayes theorem. It analyses each attribute of the data with equal importance. IBK (KNN): It is instance based classifier and it's a type of lazy learning where the function is only locally and all computation differs until classification. The experimental result analysis is showed as IBK produce a better result than the other methods for both nominal and numeric datasets. In this study class label ensemble combination methods also defined. The class label ensembles are using when the output of the combined classifiers are in the form of class labels then they can be combined using majority voting [15].

The table below shows a survey of classification algorithms we use merit and demerits.

No.	Algorithms	Advantage	Disadvantages
1	SVM	-it can be defined by convex optimization problems for which there are efficient methods. -it uses kernel trick -It has a regularization parameter which makes the user think about avoiding over-fitting. -it is an approximation to abound on the test error rate. -it able to handle both numerical and categorical variables. -is a highly accurate classifier.	-the kernel models are sensitive to over-fitting. - It lies in the choice of the kernel. -Speed and size in training and testing. Choosing the kernel is difficult.
2	IBK	-Optimized to noisy training data. - It is effective for large training data. -it handles multi-class levels. -It is easy to distance choices. -it is better to handle numeric variables.	-it needs to determine the number of the nearest neighbor. - It is difficult to distance based learning to determine clearly. -The computation cost is high. The data storage problems. -it is not good for categorical variables to handle it well.
3	J48	-Less search time. -the values generated with fewer tree approaches.	-it requires the pre-processing namely sorting.
4	NBTree	-It mitigates the effects of data loss on test attribute selections. -It is better to improve predictive accuracy	-the combined features of the decision tree and naïve Bayes classifier is somewhat complex to implement. -The pre-discritization may affect test attribute selection and decreases the predictive accuracy.

4. Proposed solutions

A deep analysis of feature selection algorithm and applying classifier ensemble methods to increase the classifier predictive accuracy in high Dimensional datasets. Basically, in this study, we aimed to conduct experimental comparisons on IBK, SVM, J48, and NBTree learning algorithms and analyze their performance with vote generalization. And using different existed feature selection algorithms namely support vector machine-recursive feature elimination (SVM-RFE), minimum redundancy maximum relevance (MrMr), info-gain, and chi-square for selecting best features depending on its usefulness or relevance criterion. Indeed the proposed approach increases the predictive accuracy of the learning algorithm.

A deep analysis of feature selection algorithm and applying classifier ensemble methods to increase the classifier predictive accuracy in high Dimensional datasets. Basically, in this study, we aimed to conduct experimental comparisons on IBK, SVM, J48, and NBTree learning algorithms and analyze their performance with vote generalization. And using different existed feature selection algorithms namely support vector machine-recursive feature elimination (SVM-RFE), minimum redundancy maximum relevance (MrMr), info-gain, and chi-square for selecting best features depending on its usefulness or relevance criterion. Indeed the proposed approach increases the predictive accuracy of the learning algorithm.

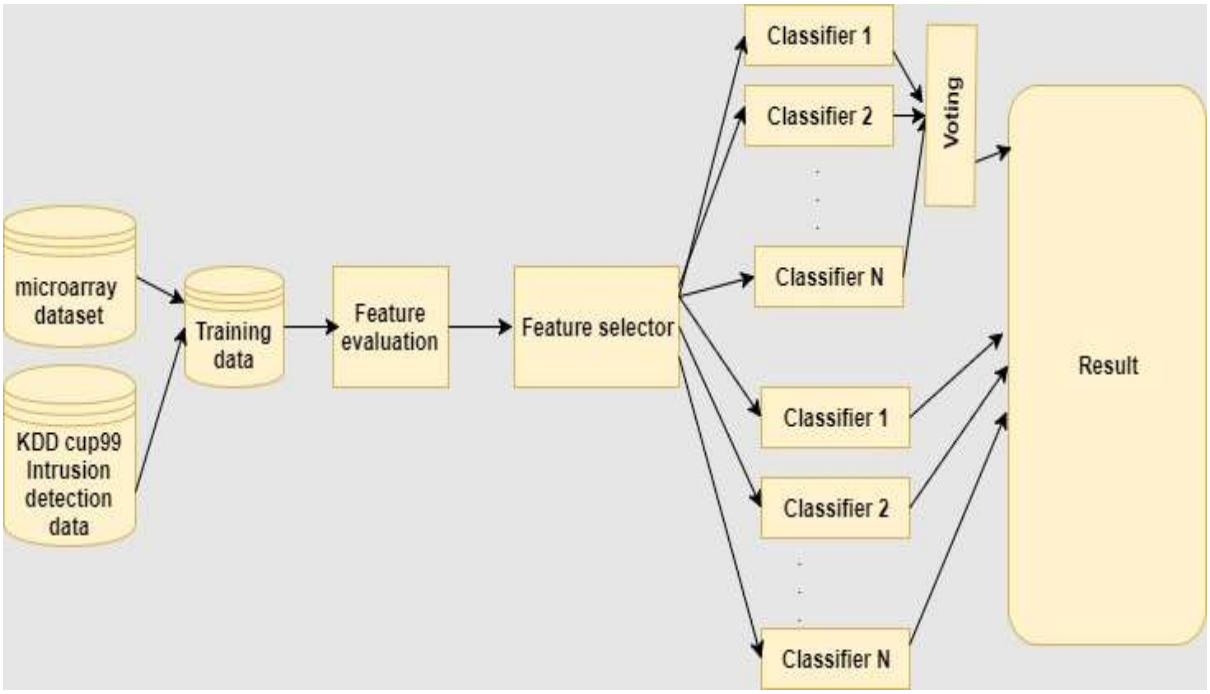

Fig1.Shows The proposed approach system architectures.

5. Conclusion

This section of feature selection algorithms and ensemble classifier of a survey shows that the feature selection algorithms always improve the stability of feature selection and a predictive accuracy of the classifier. Apart from this heterogeneous classifier is an optimized solution to model generality. Most of the time embedded and wrapper methods are computationally inefficient. Therefore developing a feature selection for high dimensional data by using a filter method is recommended. I had revised different authors ideas used in different feature selection Methods and classification algorithms that they used for solving a curse of dimensionality. In this regard, many researchers expressed as data dimensionality increases the Learning algorithm performance decreases. So it needs a Researcher to pay more attention to the analysis of feature selection algorithms with respect to the appropriate search techniques used to select the best feature subset from feature vector and classification algorithms.In the proposed approach a deep practical analysis of feature Selection algorithms and heterogenous classification with a respective feature subset search Techniques had been done. By doing these the proposed approach is an optimized solution to achieve a reduced computation time and space, unique and suitable feature subset selection, reduced

model complexity and Increased predictive model accuracy. Generally, lack of appropriate training data is a cause for the poor performance of most classifiers. So combining the output of different classifier may reduce the risk of poor performance of classifiers.

Reference

1. Aggarwal, P., & Sharma, S. K. (2015). Analysis of KDD dataset attributes-class wise for intrusion detection. Procedia Computer Science, 57, 842-851.
2. Alkuhlani, A., Nassef, M., & Farag, I. (2016). Multistage feature selection approach for high-dimensional cancer data. Soft Computing, 1-12.
3. Bolón-Canedo, V., Sánchez-Maroño, N., & Alonso-Betanzos, A. (2016). Feature selection for high-dimensional data. Progress in Artificial Intelligence, 5(2), 65-75.
4. Brahim, A. B., & Limam, M. (2017). Ensemble feature selection for high dimensional data: a new method and a comparative study. Advances in Data Analysis and Classification, 1-16.
5. Chidambaram, M., & Umasundari, R. A Survey on Feature Selection in Data Mining. ISSN: 2347-5552.
6. Dernoncourt, D., Hanczar, B., & Zucker, J. D. (2014). Analysis of feature selection stability on high dimension and small sample data. Computational Statistics & Data Analysis, 71, 681-693.
7. Destrero, A., Mosci, S., De Mol, C., Verri, A., & Odone, F. (2009). Feature selection for high-dimensional data. Computational management science, 6(1), 25-40.
8. Díaz-Uriarte, R., & De Andres, S. A. (2006). Gene selection and classification of microarray data using the random forest. BMC Bioinformatics, 7(1), 3.
9. Giancarlo, R., Bosco, G. L., & Utro, F. (2015). Bayesian versus data-driven model selection for microarray data. Natural Computing, 14(3), 393-402.
10. Gnana, D. A. A., Appavu, S., & Leavline, E. J. (2016). Literature Review on Feature Selection Methods for High-Dimensional Data. Methods, 136(1).
11. Huerta, E. B., Duval, B., & Hao, J. K. (2006, April). A hybrid GA/SVM approach for gene selection and classification of microarray data. In Workshops on Applications of Evolutionary Computation (pp. 34-44). Springer, Berlin, Heidelberg.
12. Ladha, L., & Deepa, T. (2011). Feature selection methods and algorithms. International journal of computer science and engineering, 3(5), 1787-1797.
13. Lazar, C., Taminau, J., Meganck, S., Steenhoff, D., Coletta, A., Molter, C. & Nowe, A. (2012). A survey of filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 9(4), 1106-1119.
14. Li, T., Zhang, C., & Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. Bioinformatics, 20(15), 2429-2437.
15. Nagi, S., & Bhattacharyya, D. K. (2013).Classification of microarray cancer data using ensemble approach. Network Modelling Analysis in Health Informatics and Bioinformatics, 2(3), 159-173.
16. Perthame, É. Friguet, C., & Causeur, D. (2016). Stability of feature selection in classification issues for high-dimensional correlated data. Statistics and Computing, 26(4), 783-796.
17. Rahajoe, A. D., Winarko, E., & Guritno, S. (2017). A Hybrid Method for Multivariate Time Series Feature Selection. International Journal of Computer Science and Network Security (IJCSNS), 17(3), 103.
18. Rouhi, A., & Nezamabadi-pour, H. (2017, March). A hybrid feature selection approach based on ensemble method for high-dimensional data. In Swarm Intelligence and Evolutionary Computation (CSIEC), 2017 2nd Conference on (pp. 16-20). IEEE.
19. Saleh, A. I., Talaat, F. M., & Labib, L. M. (2017). A hybrid intrusion detection system (HIDS) based on prioritized k-nearest neighbors and optimized SVM classifiers. Artificial Intelligence Review, 1-41.
20. Seijo-Pardo, B., Bolón-Canedo, V., & Alonso-Betanzos, A. (2017). Testing Different Ensemble Configurations for Feature Selection. Neural Processing Letters,1-24.
21. Vanaja, S., & Kumar, K. R. (2014). Analysis of feature selection algorithms on classification: a survey. International Journal of Computer Applications, 96(17).
22. Xing, E. P., Jordan, M. I., & Karp, R. M. (2001, June). Feature selection for high-dimensional genomic microarray data. In ICML (Vol. 1, pp. 601-608).
23. Yu, L., & Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03) (pp. 856-863).