Middle-East Journal of Scientific Research 24 (9): 2952-2956, 2016
ISSN 1990-9233

DOI: 10.5829/idosi.mejsr.2016.2952.2956

A Comparative Study on the Performance of Feature Selection Algorithms in Prediction of Down'ssyndrome in Mice Models Using Protein Expression Levels

Krithika Narayanan, Azlagiavanan Senthil, Anirruth Ragav VJ. and Shomona Gracia Jacob

Department of Computer Science, SSN College of Engineering, Chennai, India

Abstract: Data mining aims at analyzing voluminous data and extracting meaningful patterns that enable the investigator to ascertain knowledge from the data. This research work aims at exploring the data generated by measuring the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex in mice. The eight classes of mice are described based on features such as genotype, behaviour and treatment. Moreover, in order to assess the effect of the drug memantine in recovering the ability to learn in trisomic mice, some mice have been injected with the drug and others have not. The objective discussed in this research is to identify subsets of proteins that are discriminant between the classes. This is made possible by determining the feature selection algorithms that identify the best subset of features to distinguish between the classes of mice. The results of this work report that Correlation-based feature subset selection identifies the best subset of features yielding a high accuracy of 99% in predicting the class of mice based on the protein expression levels.

Key words: Selectionalgorithms · Detectablesignals · Proteinmodifications · Protein expression

INTRODUCTION

Data mining aims at analyzing voluminous data and extracting meaningful patterns that enable the investigator to ascertain knowledge from the data. In this paper, the effect of data mining techniques in predicting the type of mice based on protein expression levels is performed [1, 2]. Down's syndrome is a genetic disorder caused by the presence of an extra copy (all or in part) of chromosome 21 [3]. Memantine is currently being used as a treatment for Alzheimer's and is proposed to be used to treat Down's syndrome as well. It works by decreasing excess stimulation of a neurotransmitter that over-stimulates nerve cells causing degradation [4]. Mouse models have frequently been used to study Down's syndrome due to the close similarity in the genomes of mice and humans. In this study, the mice were given either a saline or memantine solution and they were either given a shock while in cage or roaming in room- the shock first is simulation to learn (shock-context). In the study presented in this paper, the performance of various feature selection methods on various traditionally well-performing classifiers such as Bayesian network, IBK, J48 and random forest is investigated. WEKA (Waikato Environment for Knowledge Analysis) is used to conduct the experiments. It is a non-commercial and open-source data mining system with tools for data pre-processing, classification, regression, clustering, association rules and visualization. The feature selection methods evaluated are CFS subset evaluation and Information Gain attribute evaluation.

MATERIALS AND METHODS

This section gives a brief description on the dataset used and the methods employed to determine the optimal feature subsets.

Dataset Description: The data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of cortex. There are 38 control mice and 34 trisomic mice (Down syndrome), for a total of 72 mice. In the experiments, 15 measurements were registered for each protein per sample/mouse. Therefore, for control mice, there are 38x15, or 570 measurements and for trisomic mice, there are 34x15, or 510 measurements.

The dataset contains a total of 1080 measurements per protein. Each measurement can be considered as an independent sample/mouse. The dataset is fairly

Corresponding Author: Krithika Narayanan, Department of Computer Science, SSN College of Engineering, Chennai, India.

distributed as it contains 150 instances each of class c-CS-m and class c-SC-m. There are 135 instances each of class c-CS-s, class c-SC-s, class t-CS-m, class t-SC-m, class t-SC-s and 105 instances of class t-SC-s.

Dataset Information: The eight classes of mice are described based on features such as genotype, behaviour and treatment. According to genotype, mice can be control or trisomic. According to behaviour, some mice have been stimulated to learn (context-shock) and others have not (shock-context) and in order to assess the effect of the drug memantine in recovering the ability to learn in trisomic mice, some mice have been injected with the drug and others have not.

Attribute Information:

The Attributes Given Are: Mouse ID, values of expression levels of 77 proteins, genotype, treatment type and behaviour of the mouse. The names of proteins are followed by '_n' indicating that they were measured in the nuclear fraction. For example: DYRK1A_n.

Data Pre-Processing: Initially, we remove mouse ID as it provides no useful insight towards classification. As classification should be done purely based on the protein levels, we remove genotype, treatment type and behaviour so that they don't influence the algorithms employed for classification. The missing values have been handled by averaging the existing values.

Methods: Feature selection is the process of identifying the best subset of features to predict a given set of data [5-7]. Classification is applied on the data using the selected feature sets to ascertain whether the selected features are able to yield a high prediction in accuracy [8]. Based on a survey of related work in the area of feature selection and classification, the four classifiers were identified for execution on this dataset. Bayesian Network, Nearest-Neighbour, J48, Random Forest and SMO were employed to measure the prediction accuracy using feature sets yielded by Correlation based feature subset selection and Information Gain feature selection[9].

Classification: Classification is the technique which determines to which class the data record belongs. Classification algorithms build models from the training data records given to it and this model is used to assign a class label to the new data. Random Forest and Nearest-Neighbour yielded the highest accuracy on the selected feature and hence the two algorithms are discussed in detail.

Nearest – Neighbour Instance – Based Classifier (Ibk): It is a type of lazy classification algorithm. It is a learning method in which generalization beyond the training data is delayed until a query is made to the system where the system tries to generalize the training data before receiving queries [10]. The advantage in using lazy learning is that the target function is approximated locally. But the disadvantage is that it requires large memory space to store the entire training dataset. Outliers and noisy data also have an effect on the output.

In IBK (K-nearest neighbour) classifier, the function is only approximated locally and all computation is deferred until classification. An object is classified by an approximation of its 'K' (10 in this case) nearest neighbours. 'K' is always a positive integer. Greater weights are assigned to the nearer neighbours so that they contribute more to the average than the more distant neighbours. At times, distance between neighbours could be dominated by irrelevant attributes which is overcome by elimination of the least relevant attributes in the dataset.

Table indicating details about the data set.

Data set Characteristics	Attribute Characteristics	Associated Tasks	Number of instances	Number of attributes	Missing Values
Multivariate	Real	Classification, Clustering	1080	82	Yes

UCI Mice Protein level data set with class distribution

Genotype	Behaviour	Treatment	No. of Mice	Class	Class Distribution
Control	Context-shock	Saline injection	9	c-CS-s	150
Control	Shock-context	Saline injection	9	c-SC-s	150
Control	Context-shock	Memantine injection	10	c-CS-m	135
Control	Shock-context	Memantine injection	10	c-SC-m	135
Trisomic	Context-shock	Saline injection	7	t-CS-s	135
Trisomic	Shock-context	Saline injection	9	t-SC-s	135
Trisomic	Context-shock	Memantine injection	9	t-CS-m	105
Trisomic	Shock-context	Memantine injection	9	t-SC-m	135

J48: It is an open source java implementation of C4.5 for Weka, a data mining tool developed by University of Wakaito. This algorithm is an optimized implementation of C4.5 and outputs a decision tree. Decision Trees are tools that use divide-and-conquer strategies as a form of learning by induction. It contains a root node, several intermediate nodes and leaf nodes.

Each node contains a decision and the decision leads to classification. Splitting criterion identifies the best node to split upon at the level of the tree.

Random Forest: Random Forest algorithm builds a forest (collection) of decision trees D= { hk (x, Tk) }

where
k=1,2,3…..L
L- No of decision trees
Tk-Training set built at random and identically distributed.
hk - Tree built from vector Tk and produces output x.

Trees in a Random Forest are built randomly by selecting 'm' (value fixed for all nodes) attributes in each node of the tree; where the best attribute is chosen to divide the node. The selection of a random subset of features is a type of the random subspace method, which is a way to implement the stochastic discrimination approach to classification. The vector used for training each tree is obtained using random selection of the instances. In Random Forest, to determine the class of an instance, all of the trees indicate an output 'x' (each it's own), where the most voted is selected as the final result. The classification error depends on the strength of individual trees of the forest and the correlation between any two trees in the forest is solved quickly and analytically, generally improving its scaling and computation time significantly.

Correlation Based Feature Selection: CFS (Correlation based feature subset selection) is a simple filter algorithm that ranks feature subsets according to a correlation based heuristic evaluation function [11-13]. The bias of the evaluation function is toward subsets that contain features that are highly correlated with the class and uncorrelated with each other. Irrelevant features should be ignored because they will have low correlation with the class.

The proposed methodology for classification of protein is represented below.

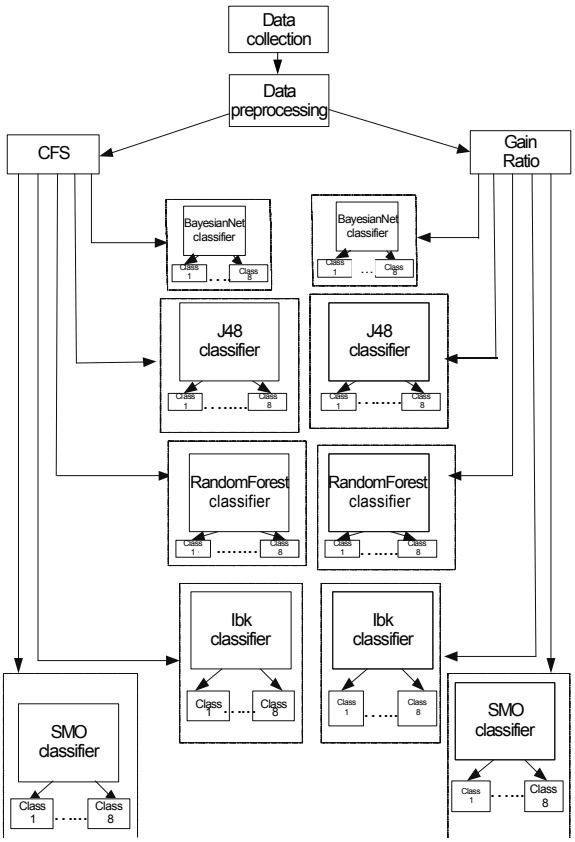

Fig: Proposed methodology for Investigation of Classifier Performance on Protein Data

The CFS algorithm has been implemented using the CfsSubsetEval(CFS Subset Evaluation) filter that evaluates the worth of the subset of attributes by considering the predictive ability of each feature individually as well as checking the degree of redundancy between them. It selects subsets containing features that have high correlation level with the class but have a very low level of inter-correlation between them[14,15].

The experimental results are discussed in the ensuing section.

EXPERIMENTAL RESULTS

The experimental results are discussed in two sections. Initially the performance of the classifiers in terms of computational time and accuracy are measured with the entire feature set used for classification. This is followed by applying the feature selection techniques to remove the irrelevant features and classifiers are implemented on the reduced optimal feature subset. The results of this experiment are tabulated in Table 1 and Table 2.

Table 1: Classifier performance prior to feature selection

Algorithms Implemented	Time Taken(Sec)	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statistic	Mean Absolute Error	Root mean square Error	Relative Absolute Error(%)	Root Relative Squared Error(%)
Bayes Net	0.35	908	172	0.8176	0.0405	0.1838	18.5158	55.6266
SMO	0.72	1059	21	0.9777	0.1877	0.2914	85.9201	88.1614
IBk	0	1077	3	0.9968	0.0025	0.0264	1.1327	7.9826
J48	0.56	945	135	0.8569	0.0344	0.175	15.7241	52.9534
Random Forest	2.1	1075	5	0.9947	0.0625	0.1174	28.5979	35.5312

Table 2: Classifier performance post feature selection (CFS)

Algorithms Implemented	Time Taken(Sec)	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statistic	Mean Absolute Error	Root mean square Error	Relative Absolute Error (%)	Root Relative Squared Error (%)
Bayes Net	0.05	922	158	0.8326	0.0428	0.1714	19.587	51.8528
SMO	0.15	985	95	0.8994	0.1885	0.2928	86.2607	88.5803
IBk	0	1077	3	0.9968	0.0025	0.0264	1.1327	7.9826
J48	0.1	913	167	0.823	0.0413	0.1922	18.8848	58.1491
Random Forest	1.37	1071	9	0.9905	0.048	0.1031	21.9501	31.1896

Table 3: Classifier performance with feature set obtained by Gain Ratio

Algorithms Implemented	Time Taken(Sec)	Correctly Classified Instances	Incorrectly Classified Instances	Kappa Statistic	Mean Absolute Error	Root mean square Error	Relative Absolute Error(%)	Root Relative Squared Error(%)
Bayes Net	0.04	849	231	0.7554	0.0605	0.2035	27.6812	61.5729
SMO	0.14	832	248	0.7369	0.1905	0.2962	87.184	89.6136
IBk	0	1059	21	0.9777	0.0066	0.0695	3.0243	21.025
J48	0.06	934	146	0.8453	0.0366	0.1774	16.7554	53.6738
Random Forest	1.12	1044	36	0.9619	0.0528	0.1186	24.189	35.8853

Table 4: Summary of Classifier Accuracy Pre- and Post- feature selection

		After Feature Selection	
Algorithm	Before Feature Selection	CFS	Information Gain
Bayes Net	84.0741%	85.3704%	78.6111%
SMO	98.0556%	91.2037%	77.037%
IBk	99.7222%	99.7222%	98.0556%
J48	87.5%	84.537%	86.4815%
Random Forest	99.537%	99.1667%	96.6667%

Accuracy is defined as the ratio of correctly classified instances to total number of instances. Accuracy was employed to compare the performance scores of the classifiers. On evaluating the models based on their accuracy and time taken to build, we found that IBk performed the best. Feature selection using gain ratio attribute evaluation was found to decrease the accuracy in all cases.With CFS subset evaluation, in the case of Bayes Net, accuracy improved by a narrow margin. In case of IBk, the accuracy remained the same. While performing feature selection based on gain ratio, a threshold of 0.5 was chosen to filter out the features. This led to a total of 12 features being selected and the accuracy obtained is tabulated below.

CONCLUSION

Data Mining is the process of extracting meaningful patterns from large datasets. This paper has investigated the performance of data mining techniques in predicting the subsets of proteins that have the ability to discriminate between the classes of mice. Moreover, the effect of feature selection has also been explored to identify if an optimal and reduced feature set is also

effective in enhancing the classifier accuracy. However it is observed that the classifier accuracy is diminished when reduced number of features is employed for classification. Hence from this it is ascertained that better feature selection algorithms are required to identify the optimal set of features for prediction of protein subsets.

REFERENCES

1. Jacob, S.G. and R.G. Ramani, 2015. Prediction of Rescue Mutants to Restore Functional Activity of Tumor Protein TP53 through Data Mining Techniques, Journal of Scientific & Industrial Research, 74: 135-140.
2. Geetha Ramani, R. and S.Gracia Jacob, 2013. Prediction of cancer rescue p53 mutants in silico using Naïve Bayes learning methodology, Protein and peptide letters, 20(11): 1280-1291.
3. Higuera, C., K.J. Gardiner and K.J. Cios, 2015. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome, PloS one, 10(6): 129126.
4. Ahmed, M.M., A.R. Dhanasekaran, A. Block, S. Tong, A.C. Costa, M. Stasko and K.J. Gardiner, 2015. Protein dynamics associated with failed and rescued learning in the Ts65Dn mouse model of Down syndrome. PloS one, 10(3): 119-491.
5. Kuang, Y., 2009. A Comparative Study on Feature Selection Methods and Their Applications in Causal Inference, Institutionen for datavetenskap, Lunds universitet.
6. Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification, Journal of Machine Learning Research, 3: 1289-1305.
7. Li, X., X. Gong, X. Peng and S. Peng, 2013. SSiCP: a new SVM based recursive feature elimination algorithm for multiclass cancer classification, Bio-Medical Materials and Engineering, 23: 1027-1038.
8. Li, T., C. Zhang and M. Ogihara, 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression, Bioinformatics, 20(15): 2429-2437.
9. Ramani, R.G. and S.G. Jacob, 2013. Benchmarking classification models for cancer prediction from gene expression data: A novel approach and new findings, Studies Informatics Control, 22(2): 134-143.
10. Jacob, S.G. and R.G. Ramani, 2013. Design and Implementation of a clinical data classifier: A Supervised learning approach, Res. J. Biotech, 8(2): 16-26.
11. Abusamra, H., 2013. A comparative study of feature selection and classification methods for gene expression data of glioma, Procedia Computer Science, 23: 5-14.
12. Ramani, R.G. and S.G. Jacob, 2013. Prediction of P53 mutants (multiple sites) transcriptional activity based on structural (2D&3D) properties, PloS one, 8(2): 55401.
13. Ramani, R.G. and S.G. Jacob, 2013. Improved classification of lung cancer tumors Based on structural and physicochemical properties of proteins using data mining models, PloS one, 8(3): 587-72.
14. Dunham, M.H., 2006. Data mining: Introductory and advanced topics, Pearson Education India.
15. Jacob, S.G. and R.G. Ramani, 2013. Design and Implementation of a clinical data classifier: A Supervised learning approach, Res. J. Biotech, 8(2): 16-26.