Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Technology 1 (2012) 323 – 327

Procedia Technology

INSODE 2011

A comparative study on the effect of feature selection on classification accuracy

Esra Mahsereci Karabulut[a*], Selma Ayşe Özel[b], Turgay İbrikçi[b,c]

[a] Gaziantep University, Gaziantep Vocational High School, Şehitkamil/Gaziantep, 27310, Turkey
[b] Çukurova University, Electrical-Electronics Engineering Department, Balcalı/Adana, 01330, Turkey
[c] Çukurova University,Computer Engineering Department, Balcalı/Adana, 01330, Turkey

Abstract

Feature selection has become interest to many research areas which deal with machine learning and data mining, because it provides the classifiers to be fast, cost-effective, and more accurate. In this paper the effect of feature selection on the accuracy of NaïveBayes, Artificial Neural Network as Multilayer Perceptron, and J48 decision tree classifiers is presented. These classifiers are compared with fifteen real datasets which are pre-processed with feature selection methods. Up to 15.55% improvement in classification accuracy is observed, and Multilayer Perceptron appears to be the most sensitive classifier to feature selection.

Keywords: feature selection; classification; accuracy; Naïve Bayes; multilayer perceptron

1. Introduction

Feature selection is the process of removing redundant or irrelevant features from the original data set. So the execution time of the classifier that processes the data reduces, also accuracy increases because irrelevant features can include noisy data affecting the classification accuracy negatively [1]. With feature selection the understandability can be improved and cost of data handling becomes smaller [2].

Feature selection algorithms are divided into three categories; filters, wrappers and embedded selectors. Filters evaluate each feature independent from the classifier, rank the features after evaluation and take the superior ones [3]. This evaluation may be done using entropy for example [4]. Wrappers takes a subset of the feature set, evaluates the classifier's performance on this subset, and then another subset is evaluated on the classifier. The subset for which the classifier has the maximum performance is selected. So wrappers are dependent on the selected classifier. In fact wrappers are more reliable because classification algorithm affects the accuracy, although the selection of the subset is an NP-hard problem [5]. So it takes considerable processing time and memory. Some heuristic algorithms can be used for subset selection such as genetic algorithm, greedy stepwise, best first or random search. Therefore, the filters are more time efficient when compared to wrappers, but they don't take into account that selecting the

* Esra Mahsereci Karabulut. Tel.: +0-506-882-7527; fax: +0-342-221-2412.
E-mail address: mahsereci@gantep.edu.tr.

doi:10.1016/j.protcy.2012.02.068

better features may relevant to classification algorithms [6]. Embedded techniques on the other hand perform feature selection during learning process like artificial neural networks do.

In literature there are many feature selection studies including one on why to prefer filters to wrappers [7] and feature selection methods for classification [8, 9]. In this study, we used fifteen datasets to compare three classification algorithms with respect to their influence from six feature selection filters.

In Sections 2 and 3 algorithms and datasets used in this study are presented respectively. In Section 4, the classification accuracies are compared. Conclusions are given in the last section.

2. Filter and Classification Algorithms

In filter algorithms, features are first scored and ranked according to the relevance to the class label, and then are selected according to a threshold value [10]. Each of these feature selection algorithms has an evaluation value for each feature. Features having an evaluation value greater than the threshold are selected. Threshold is predefined by user, in this study it is defined as 0.1. Filter algorithms used for feature selection are; Information Gain [11], Gain Ratio [12], Symmetrical Uncertainty [13], Relief-F [14], One-R [15] and Chi-square [16].

A classification algorithm assigns instances to a category according to a given set of features [17]. When classification is performed on the output of a feature selection process, the prediction becomes more accurate and time efficient. In this study, classification algorithms used with feature selection are Naïve Bayes [18], Multilayer Perceptron (MLP) [19] and J48 decision tree [20].

3. Data Sets

For experiments, fifteen data sets are taken from Data Mining Repository of University of California Irvine (UCI) [21]. A brief summary of these datasets are given in Table1.

Table 1. Data Sets

Dataset	# of examples	# of features	Type	# of classes
Audiology	226	69	Discrete	24
Balance-scale	625	4	Discrete	3
Breast-cancer	286	9	Mixed	2
Car	1728	6	Discrete	4
Credit	690	15	Mixed	67
Ionosphere	351	32	Continuous	2
Iris	150	4	Continuous	3
Lung-cancer	32	56	Discrete	3
Lymphography	148	18	Discrete	4
Mushrooms	8416	22	Discrete	2
Post-operative	90	8	Mixed	3
Primary-tumor	339	17	Discrete	21
Vehicle	946	18	Continuous	4
Vowel	990	10	Continuous	11
Zoo	101	16	Discrete	7

4. Experimental Results

To compare the performance of the classification algorithms with feature selection methods, WEKA data mining tool was used, the default parameters were used for each classification algorithm [22]. All experiments were carried out using a ten-fold cross validation approach.

Five feature selection algorithms are employed to select features before passing the data sets to the classifiers. Threshold value is selected as 0.1 to select the features. Classification accuracies are presented before and after the feature selection in Table2, Table3 and Table4. The columns named Full presents the accuracy values of classification using all features, i.e no feature selection. The bold written values indicate the increase in accuracy comparing to Full value of the data set.

Table 2. Effects of feature selection on Naïve Bayes

Dataset	Classification Accuracy (%)						
	Full	Inf. Gain	Gain Ratio	Symmetrical Uncertainty	Relief-F	One-R	Chi-square
Audiology	73.45	**73.89**	73.00	67.70	68.14	73.45	73.00
Balance-scale	90.40	88.96	50.56	47.68	57.76	90.40	90.40
Breast-cancer	71.68	70.28	70.28	70.28	66.78	71.68	71.68
Car	85.53	76.85	76.85	76.85	83.10	85.53	85.53
Credit	77.68	76.96	74.78	74.78	**84.93**	77.68	77.68
Ionosphere	82.62	**83.19**	**83.48**	**83.48**	82.62	82.62	82.62
Iris	96.00	96.00	96.00	96.00	96.00	96.00	96.00
Lung-cancer	78.12	75.00	78.12	78.12	75.00	78.12	75.00
Lymphography	83.11	80.40	77.70	77.70	77.02	83.11	83.11
Mushrooms	95.83	95.72	**96.32**	**96.13**	-	95.83	95.83
Post-operative	66.67	**71.11**	**71.11**	**71.11**	**71.11**	66.67	66.67
Primary-tumor	50.15	49.85	46.90	36.58	42.48	50.15	50.15
Vehicle	44.80	42.67	**45.27**	43.14	25.65	44.80	44.80
Vowel	63.74	**67.88**	**67.68**	62.73	56.06	**63.94**	63.54
Zoo	95.05	95.05	95.05	**96.04**	**96.04**	95.05	95.05

According to Table 2 Naïve Bayes classifier is mostly affected by Gain Ratio at five data sets, but also for the Credit dataset accuracy is increased from 77.68 to 84.93 as the largest difference. Chi-square feature selection algorithm doesn’t affect Naïve Bayes positively with respect to these datasets. Eight of fifteen datasets are affected from at least one feature selection algorithm.

Table 3. Effects of feature selection on MLP

Dataset	Classification Accuracy (%)						
	Full	Inf. Gain	Gain Ratio	Symmetrical Uncertainty	Relief-F	One-R	Chi-square
Audiology	83.19	79.65	82.74	72.12	72.57	**83.63**	82.74
Balance-scale	90.72	89.76	50.40	47.68	58.24	90.56	**90.88**
Breast-cancer	64.69	**70.28**	**70.28**	**70.28**	**66.78**	**65.38**	**68.53**
Car	99.54	77.43	77.43	77.43	86.52	**99.65**	99.30
Credit	84.20	83.19	83.19	82.89	82.46	83.91	82.03
Ionosphere	91.17	**92.02**	**91.45**	**92.02**	84.05	91.17	**91.74**
Iris	97.33	**98.00**	97.33	97.33	98.00	97.33	98.00
Lung-cancer	65.63	**75.00**	**78.13**	**78.13**	**75.00**	65.63	**68.75**
Lymphography	84.46	79.73	80.41	81.08	79.73	83.11	84.46
Mushrooms	-	-	-	-	-	-	-
Post-operative	55.56	**71.11**	**71.11**	**71.11**	**71.11**	**60.00**	**57.78**
Primary-tumor	38.35	**41.59**	**42.77**	**40.71**	37.46	**41.59**	**40.12**
Vehicle	81.68	80.61	76.95	74.47	25.65	**82.15**	81.32
Vowel	92.73	79.39	79.29	67.78	63.43	92.02	**92.93**
Zoo	96.04	96.04	96.04	96.04	96.04	96.04	96.04

Table 4: Effects of feature selection on decision tree J48

Dataset	Classification Accuracy (%)						
	Full	Inf. Gain	Gain Ratio	Symmetrical Uncertainty	Relief-F	One-R	Chi-square
Audiology	77.88	**78.32**	**78.32**	72.12	74.34	**78.32**	**78.32**
Balance-scale	76.64	76.00	50.56	47.68	56.80	76.64	76.64
Breast-cancer	75.52	70.28	70.28	70.28	66.78	75.52	75.52
Car	92.36	92.36	76.56	76.56	86.34	92.36	92.36
Credit	86.09	85.07	85.07	85.51	85.07	86.09	**86.23**
Ionosphere	91.45	91.17	91.45	91.45	87.46	91.17	91.17
Iris	96.00	96.00	96.00	96.00	96.00	96.00	96.00
Lung-cancer	78.13	78.13	78.13	78.13	75	78.13	78.13
Lymphography	78.38	75.00	77.03	77.03	77.70	77.03	76.35
Mushrooms	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Post-operative	70.00	**71.11**	**71.11**	**71.11**	**71.11**	70.00	70.00
Primary-tumor	39.82	38.64	38.64	37.46	38.94	39.82	39.82
Vehicle	72.46	**74.11**	**73.88**	69.86	25.65	**72.58**	72.46
Vowel	81.52	**82.02**	81.41	78.28	70.51	81.21	81.21
Zoo	92.08	91.09	90.01	**94.06**	**93.07**	90.10	92.08

According to Table 3, MLP is more affected by feature selection algorithms than Naïve Bayes. Ten of the datasets are affected by at least one feature selection algorithm positively. MLP is mostly affected by Chi-square in positively affected seven datasets. Post-operative is the most affected dataset, the increase in accuracy is 15.55 %, from 55.56% to 71.11% as can be seen in Table3.

Table 4 indicates that J48 decision tree classifier is not affected by feature selection as much as Naïve Bayes and MLP. Six of the datasets have an increase in accuracy, but the maximum increase is 1.98%. Information Gain has the positive affect on four datasets outperforming others for J48.

5. Conclusions

Feature selection is an important issue in classification, because it may have a considerable effect on accuracy of the classifier. It reduces the number of dimensions of the dataset, so the processor and memory usage reduce; the data becomes more comprehensible and easier to study on.

In this study we have investigated the influence of feature selection on three classifiers Naïve Bayes, MLP and decision tree J48 using fifteen real-life datasets. We observed that MLP is the most affected classifier; ten of the used datasets are more accurately classified by preprocessing of at least one feature selector. The classification accuracy is improved up to 15.55% in Post-operative dataset. It is also observed for Naïve Bayes classifier the Gain Ratio, for MLP the Chi-square and for J48 the Information Gain is the most positively effective feature selection algorithm.

References

1. S. Doraisami, S. Golzari, A Study on Feature Selection and Classification Techniques for Automatic Genre Classification of Traditional Malay Music, Content-Based Retrieval, Categorization and Similarity, 2008
2. A. Arauzo-Azofra, J. L. Aznarte, and J. M. Benítez, Empirical study of feature selection methods based on individual feature evaluation for classification problems, Expert Systems with Applications, 38 (2011) 8170-8177.
3. I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res., 3 (2003) 1157-1182.
4. Y. Özkan., “Veri Madenciliği Yöntemleri”, Papatya Publication, İstanbul, 2008
5. J. Novakovic, The Impact of Feature Selection on the Accuracy of Naive Bayes Classifier, 18th Telecommunications forum TELFOR, 2010
6. M. A. Hall and L. A. Smith, Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper, Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference, 1999.
7. G. John, R. Kohavi, and K. Pfleger, Irrelevant Features and the Subset Selection Problem, International Conference on Machine Learning, (1994)121-129.
8. M. Dash and H. Liu, Feature selection for classification, Intelligent Data Analysis, 1(3) (1997) 131–156.
9. J. Brank, M. Grobelnik, N. Milic-Frayling and D. Mladenic, Interaction of Feature Selection Methods and Linear Classification Models, Proceedings of the ICML-02 Workshop on Text Learning, 2002.
10. I. H. Witten and M. A. Hall, Data mining : practical machine learning tools and techniques, Amsterdam; Boston, Morgan Kaufmann, 2011.
11. T. M. Cover and J. A. Thomas, Elements of information theory, 2nd ed. Hoboken, N.J.: Wiley-Interscience, 2006.
12. T. M. Mitchell, Machine learning, Boston, WCB/McGraw-Hill, 1997.
13. W. H. Press, Numerical recipes in C : the art of scientific computing, 2nd ed. New York, 1988.
14. K. Kira and L. A. Rendell, A practical approach to feature selection, Proceedings of the ninth international workshop on Machine learning, Aberdeen, Scotland, United Kingdom, 1992.
15. R. Holte, Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, Machine Learning, 11 (1993) 63-90.
16. H. Liu and R. Setiono, Chi2: Feature Selection and Discretization of Numeric Attributes, Proceedings of the Seventh International Conference on Tools with Artificial Intelligence, 1995.
17. J. Han and M. Kamber, Data mining : Concepts and Techniques, Amsterdam; Boston, Morgan Kaufmann, 2006.
18. G. John and P. Langley, Estimating Continuous Distributions in Bayesian Classifiers, Eleventh Conference on Uncertainty in Artificial Intelligence, (1995) 338-345.
19. F. Rosenblatt, Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms, Spartan, 1962.
20. R. Quinlan, C4.5: Programs for Machine Learning (Morgan Kaufmann Series in Machine Learning): Morgan Kaufmann, San Mateo, 1992.
21. D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, UCI Repository of machine learning databases, University California Irvine, Department of Information and Computer Science, 1998
22. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten, The WEKA data mining software: an update, SIGKDD Explor. Newsl., 11 (2009) 10-18.