IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 2, February 2016.
www.ijiset.com
ISSN 2348 – 7968

A Novel Approach of Feature Selection Techniques for Image Dataset

A. Sorna Gowri
Assistant Professor, The M.D.T Hindu College, Tirunelveli
gowrimurugan74@yahoo.co.in

Dr. K. Ramar
Principal, Einstein college of Engineering, Tirunelveli
Kramar.einstein@gmail.com

Abstract

Feature selection techniques have become an obvious need for researchers in computer science and many fields of science. Whether the target research is in medicine, agriculture, business, or industry, the necessity for analyzing large amount of data is needed. Addition to that, finding the most excellent feature selection technique that best satisfy a certain learning algorithm could bring the benefit for the research and researchers. Therefore, a new method has been proposed for diagnosing breast cancer on a combination of learning algorithm tools and features selections techniques. The idea is to obtain a hybrid approach that combines between the best performing learning algorithms and the best performing features selections techniques. The experiment result shows that assemblage between correlation based features selections method along with Naïve Bayes learning algorithm can produce a promising results. However, no single feature selection methods that best satisfy all datasets and learning algorithms. Therefore, machine learning researchers should understand the nature of datasets and learning algorithms characteristics in order to obtain better outcomes.

Keywords: WBC, K-NN, Naïve bayes, Decision Tree

Introduction:

The advancement of information technology, the growing number of social networks websites, electronic health information systems, and other factors have flooded internet with data. The amount of data posted daily on internet is increasing daily. At the same time, not all data are important or even needed. Therefore, data mining researches started using the term features selections or data selections more often. Feature selection or attribute subset combination is the process of identifying and utilizing the most relevant attributes and removing as many redundant and irrelevant attributes as possible [2]. In addition, features selections mechanisms do not alter the original representation of data in any way. It just selects an optimal useful subset. Recently, the inspiration for applying features selection techniques in machine learning has shifted from theoretical approach to one of steps in model building. Many attribute selection methods use the task as a search problem, where each result in the search space groups a distinct subset of the possible attributes. Since the space is exponential in the number of attributes which produce lots of possible subsets, this requires the use of a heuristic search procedure for all data sets. The search procedure is combined with an attribute utility estimator in order to evaluate the relative merit of alternative subsets of attributes .This large number of possible subsets and the computation cost involved necessitate researchers to conduct a benchmark feature selection methods that produce the best possible subset in regards to more accurate results as well as low computation overhead. Feature selection techniques could perform better if the researcher chooses the right learning

ISSN 2348 – 7968

algorithm. Therefore, a new approach proposed which combines a promising feature selection technique and one of well-known learning algorithm. In the current work, we have focused on publicly available diseases datasets (Breast cancer) to evaluate the proposed approach.

II. Feature Selection Techniques

The literature showed many methods for selecting subset of features concentrate in Correlation based Feature Selection (CFS), Information Gain (IG), Relief (R), Principle Components Analysis (PCA), Consistency based Subset Evaluation (CSE), and symmetrical uncertainty (SU). CFS aims to find subsets that contain features that are highly correlated with the class and uncorrelated with each other [5]. IG is one of the simplest attribute ranking methods that rank the quality of attribute according to the difference between prior and post entropy [6]. R objective is to measure the quality of attributes according to how their values distinguish instances of different classes [3]. PCA is probably the oldest feature selection method. It is aim is to reduce the dimensionality of a data set in which there are a large number of correlated features and keeping the uncorrelated features present in the data set [4]. CSE try to obtain a set of attributes that divide the original dataset into subsets that contain one class majority , while SU is a modified information gain method that compensate the information gain bias [2].

III .The Experiment Methodology

Different sets of experiments were performed to evaluate benchmark attributes selections methods on well-known publicly available dataset from UCI machine learning repository, Wisconsin Breast Cancer dataset (WBC) [1]. For obtaining a fair judgment, as possible, between feature selection methods, this work considered three machine learning algorithms from three categories of learning methods. The first algorithm is *k*-nearest neighbors (*k*-NN) from lazy learning category. *k*-NN is an instance-based classifier where the class of a test instance is based upon the class of those training instances alike to it. Distance functions are common to find the similarity between instances. Examples of distance functions are Euclidean and Manhattan distance functions .The second algorithm is Naïve Bayes classifier (NB) from Bayes category. NB is a simple probabilistic classifier based on applying Bayes' theorem. NB is one of the most efficient and effective learning algorithms for machine learning and data mining because the condition of independency (no attributes depend on each other) [7]. The last machine learning algorithm is Random Tree (RT) or classification tree. RT is used to classify an instance to a predefined set of classes based on their attributes values. RT is frequently used in many fields such as engineering, marketing, and medicine . After applying features selections techniques and the learning algorithms on the dataset and obtaining classification accuracy results, a hybrid method will be constructed that combines the advantages of best performed feature selection technique and the advantages of best perform learning algorithm as shown in Figure1.

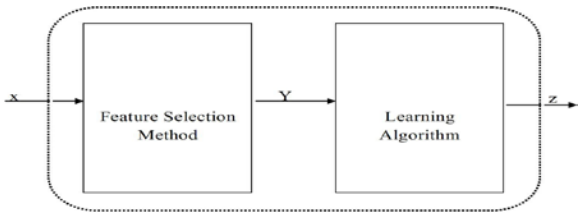

Figure1: Hybrid method of feature selection technique and a learning algorithm

The software package used in the present paper is Waikato Environment for Knowledge Analysis (WEKA). Weka

ISSN 2348 – 7968

provides the environment to perform many machine learning algorithm and feature selection methods. Weka is an open source machine learning software written in JAVA language. WEKA contains some data mining and machine learning methods for data pre-processing, classification, regression, clustering, association rules, and visualization [8].

IV. The Experimental Results

The notations "+", "-", and "=" are used to show the feature selection methods classification performance in compared with the original dataset (before performing feature selection methods); where "+" denotes to improvement, "-" denotes to degradation, and "="denotes unchanged. The experimental results of using Naïve Bayes (NB) as a machine learning algorithm on WBC dataset is shown in Table 1.

Table 1: Results for Attributes Selection Methods with Naïve Bayes.

Method	WBC
Original Dataset	95.99%
Correlation based feature selection(CFS)	95.99%=
Information Gain (IG)	95.99%=
Relief (R)	95.99%=
Principle Components Analysis (PCA)	96.14%+
Consistency based Subset Evaluation (CSE)	96.28%+
Symmetrical Uncertainty (SU)	95.99%=

Table 1 shows the results of applying WBC dataset on Naïve Bayes learning method and some features selections techniques. It showed that classification accuracy of using Naïve Bayes on original WBC dataset is 95.99%, where it showed improvement by applying features selections methods Principle Components Analysis and Consistency bases Subset Evaluation. The best result was performed by Consistency bases Subset Evaluation technique about 96.28% of classification accuracy, while classification accuracy stayed the same by using correlation based feature selection, information gain, Relief, and Symmetrical Uncertainty. Figure 2 illustrates the results on Table 1.

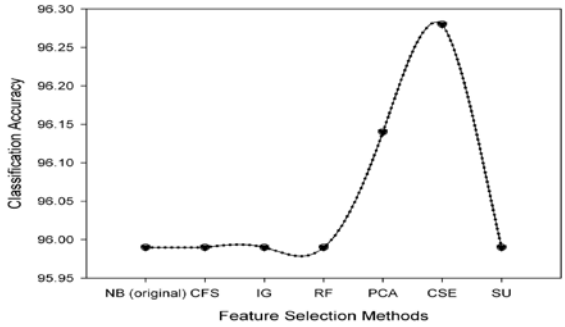

Figure2: Attributes selection methods with Naïve Bayes

The second machine learning classifier for testing features selections methods is *k*-NN. The experimental results of using *k*-NN as a machine learning algorithm on WBC is shown in Table2.

Table2: Results for Attributes Selection Methods with *k*-NN

Method	WBC
Original Dataset	95.42%
Correlation based feature selection(CFS)	95.42%=
Information Gain (IG)	95.42%=
Relief (R)	95.42%=
Principle Components Analysis (PCA)	95.42%=
Consistency based Subset Evaluation (CSE)	95.85%+
Symmetrical Uncertainty (SU)	95.42%=

Table 2 shows that the classification accuracy of using *k*-NN on the original WBC is 95.42%, where it shows

ISSN 2348 – 7968

improvement by applying the features selections method Consistency based Subset Evaluation (CSE). On the other hand, other features selections methods produced the same classification accuracy as the original dataset. Figure 3 illustrates the results on Table2.

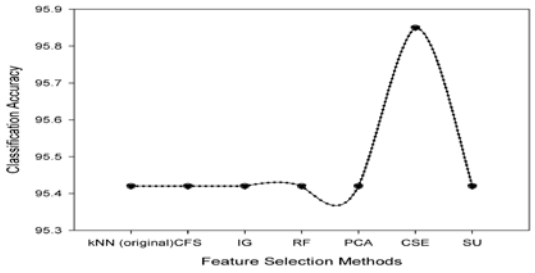

Figure3: Results for attributes selection methods with *k*-NN

The last machine learning classifier in our experiment is Decision Tree (DT). The experimental results of using DT as a machine learning algorithm on WBC is shown in Table 3

Table3: Results for Attributes Selection Methods with Decision Tree

Method	**WBC**
Original Dataset	94.56%
Correlation based feature selection(CFS)	94.56%=
Information Gain (IG)	94.56%=
Relief (R)	94.56%=
Principle Components Analysis (PCA)	94.85%+
Consistency based Subset Evaluation (CSE)	93.56%-
Symmetrical Uncertainty (SU)	94.56%=

In Table 3 shows improvement in classification accuracy by applying the features selections PCA. There is a decline in classification accuracy by using CSE, where the classification accuracy is not changed using CFS, IG, R, and SU. Figure 4 illustrates the results on Table 3.

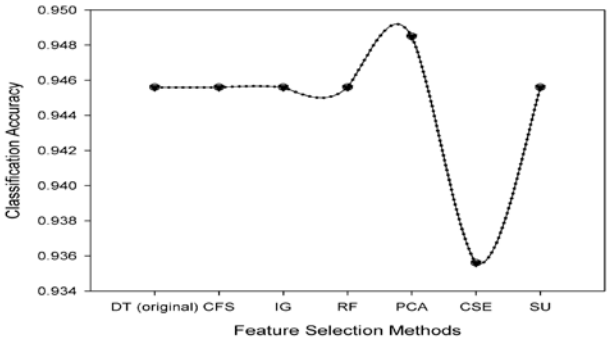

Figure 4: Results for attributes selection methods with Decision Tree

V. Conclusion:

According to the results obtained by the current work on WBC, Naïve Bayes has performed the supreme in regard to classification accuracy. *K-NN* and DT have performed just better on dataset after applying features selections methods. In general, features selections methods can improve the performance of learning algorithms. However, no single features selections method that best satisfy all datasets and learning algorithms. Therefore, machine learning researcher should understand the nature of datasets and learning algorithm characteristics in order to obtain better outcomes as possible. Overall, CSE features selection method performed better than IG, SU, R, CFS, and PC. The study also found that IG and SU performed typically because SU is a modified version of IG.

References

1. Wolberg, W. and L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences*, 1990. 87: p. 9193 - 9196.
2. Rutkowski, L., et al., eds. Artificial Intelligence and Soft Computing, Part I.

ed. L.N.i.C.S. 6113. Vol. 1. 2010, *Springer*: Poland. 487-498.
3. Guyon, I. and A. Elisseeff, An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 2003. 3: p. 1157-1182.
4. Liu, H. and R. Setiono. A probabilistic approach to feature selection: A fiter solution. *in Proceedings of the 13th International Conference on Machine Learning*. 1996. Morgan Kaufmann.
5. Hall, M.A. and L.A. Smith, Feature subset selection: a correlation based filter approach. 1997.
6. Forman, G., An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 2003. 3: p. 1289-1305.
7. Zhang, H. and J. Su, Naïve Bayes for optimal ranking. *Journal of Experimental & Theoretical Artificial Intelligence*, 2008. 20(2): p. 79-93.
8. Ashraf, M., K. Le, and X. Huang, Information Gain and Adaptive Neuro-Fuzzy Inference System for Breast Cancer Diagnoses, *in International Conference on Computer Sciences and Convergence Information Technology (ICCIT)*. 2010, IEEE: Seoul. p. 911-915.
9. Buxton, B.F., W.B. Langdon, and S.J. Barrett, Data Fusion by Intelligent Classifier Combination. *Measurement and Control*, 2001. 34(8): p. 229-234.
10. Goonatilake, S. and S. Khebbal, *Intelligent Hybrid Systems*. 1994: John Wiley & Sons, Inc.
11. Tsoumakas, G., L. Angelis, and I. Vlahavas, Selective fusion of heterogeneous classifiers. *Intelligent Data Analysis*, 2005. 9(6): p. 511-525.
12. Džeroski, S. and B. Ženko, Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 2004. 54(3): p. 255-273.
13. Kuncheva, L. and C. Whitaker, Feature Subsets for Classifier Combination: An Enumerative Experiment, in Multiple Classifier Systems, J. Kittler and F. Roli, (Editors). 2001, *Springer Berlin Heidelberg*. p. 228-237.
14. Jurman, G., Riccadonna, S., Visintainer, R., & Furlanello, C., Canberra distance on ranked lists. *In Proceedings, Advances in Ranking–NIPS 09 Workshop* , 2009, p. 22-27.