Interval-based Attribute Evaluation Algorithm

Mostafa A. Salama[1], Nashwa El-Bendary[2] Aboul Ella Hassanien[3], Kenneth Revett[1], Aly A. Fahmy[3]

[1]*Department of Computer Science, British University in Egypt, Cairo, Egypt*
Email: mostafa.salama, ken.revett@bue.edu.eg
[2] *Arab Academy for Science, Technology, and Maritime Transport, Cairo, Egypt*
Email: nashwa m@aast.edu
[3]*Cairo University, Faculty of Computers and Information, Cairo, Egypt*
Email:aboitcairo, aly.fahmy@gmail.com

***Abstract*—Attribute values may be either discrete or continuous. Attribute selection methods for continuous attributes had to be preceded by a discretization method to act properly. The resulted accuracy or correctness has a great dependance on the discretization method. However, this paper proposes an attribute selection and ranking method without introducing such technique. The proposed algorithm depends on a hypothesis that the decrease of the overlapped interval of values for every class label indicates the increase of the importance of such attribute. Such hypothesis were proved by comparing the results of the proposed algorithm to other attribute selection algorithms. The comparison between different attribute selection algorithms is based on the characteristics of relevant and irrelevant attributes and their effect on the classification performance. The results shows that the proposed attribute selection algorithm leads to a better classification performance than other methods. The test is applied on medical data sets that represent a real life continuous data sets.**

***Index Terms*—Attribute selection; Classification, ChiMerge.**

I. Introduction

ONE OF the major problems in data mining tools is the curse of dimensionality, several attribute reduction algorithms have been developed to solve such problem. The high number of attributes may contain irrelevant or redundant attributes to the classification methods [1]. Attribute reduction algorithms are either attribute selection or attribute extraction algorithms. Attribute selection algorithms determine the importance of the attributes according to the class labels. The first selected attribute that got the highest rank is the most relevant attribute to the class labels, then the relevance degree decreases until the least ranked attribute. If the classifier is applied only on attributes of the highest ranked attributes, the accuracy of the classifier should be better than being applied on all attributes. The reason of the decrease in accuracy when using all attributes is that the attributes with the lowest ranks have a negative impact on the classification result. An interesting observation that appears in [2], [3] that the trend of the classification accuracy of the classifier applied after the attribute selection algorithm increases until a certain peak where the most relevant attributes are used attributes.Then the classification accuracy starts to decrease which shows the effect of the irrelevant attributes, attributes with the lowest ranks, on the classifier.

The input data sets are either contain attributes of continuous values, discrete values or both types. For continuous data sets, attribute selection algorithms like chi-square, gain ration and information gain have to be preceded by a discretization method [4]. The correctness of the selected attributes has a great dependence on such discretization method. The proposed method here is an attribute selection and ranking method of continuous attributes that does not need to be preceded by a discretization method. It depends on a hypothesis that as the non-overlapped interval of values between the classes labels of an attribute increases as the importance of this attribute increases. This algorithm calculates the number of values in these non-overlapped interval for each attribute and accordingly creates a ranking vector. The proposed algorithm will be compared to other algorithms through checking which attribute selection algorithm would lead to the maximum classification accuracy with the least number of attributes. And hence, two numbers will be used in the comparison which are the number of attributes of highest classification accuracy and the value of this accuracy.

The proposed algorithm will be applied on two different medical real life data sets which are the Indian diabetes and HCV data sets. The classification methods used are the Multi-layered Perceptron and Support vector machine implemented by Weka software.

The rest of this paper is organized as follows: Section II shows the proposed interval-based attribute selection algorithm. Classification results and comparisons with different attribute selection algorithms illustrated in section III. Conclusion and future work is discussed in section IV.

II. The Interval-based Attribute Selection Algorithm

The interval-based attribute selection algorithm depends on a hypothesis that as the intersection between attribute value ranges of different class labels decreases as the importance of this attribute increases. The reason of this hypothesis is that if an attribute has a certain continuous range of values appears only in the case of a certain class label, then this attribute can help as an indication to this class label. Moreover, as the length of this kind of ranges increases as the importance of this

978-83-60810-22-4/$25.00 © 2011 IEEE

attribute increases. Figure (1) shows an attribute that contains an interval for each class.

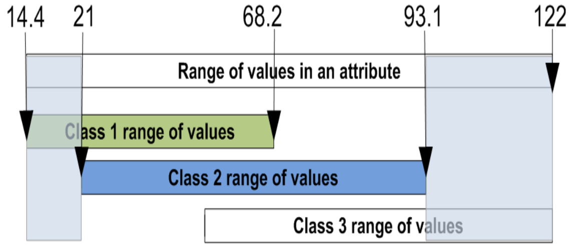

Fig. 1. Non overlapping intervals

The dashed areas show the ranges of values that are not overlapped between multiple class, only a single class label is assigned to this label. In order to evaluate the importance of such attribute, the number of values in ranges that falls in a single class will be calculated for every class and summed. i.e. as shown in figure 1, the number of values that falls in the dashed areas are counted. Then the resulted value, after refinement of this count as shown in the equation 5, will be considered as the attribute rank among other attributes.

$$\mu_a = \frac{1}{n} * \sum_{c \in C} \frac{n_{ci}}{n_c} \tag{1}$$

μ_a represents the rank of attribute a, n is the number of objects in the data set, n_c is the number of values where the corresponding objects are of class label c, and n_{ci} is the number of values in a rang that is completely falls in class c where this range is not overlapped with other class labels. For a two class data set, algorithm(1) can be used to calculate the interval-based ranking value which is μ_a. The algorithm detects, for every class, the range of values for an attribute that are not in the class and hence counts the number of objects in that range.

The removal of misleading values in Algorithm(2) is an optional step as it depends on the collection methodologies whether it is accurate or not. This step decrease the sensitivity to outliers by removing the values that are most far away from the average of the attribute values. This step should remove only a small percentage of the values in the attribute in order not to affect the accuracy of the results.

Algorithm 1 Calculate Interval-based rank μ_a of attribute a

```
$\mu_a$ : Attribute $a$'s rank, initial value is 0
$AttributeLength$ : Number of objects
$x_a$ and $n_a$ : max and min values of attribute $a$
for Each Class label $c$ do
  Remove misleading values.
  Determine the interval that represent the range of values of the attribute in that class label.
  $IntervalLength$ : Number of objects in class $c$
  $x_{ac}$ and $n_{ac}$ : The max and min values of this interval .
  //Calculate the number of values outside the interval range.
  $\mu_c$ : Initial value is 0
  for Each value $v$ in attribute $a$ do
    if $v < n_{ac}$ or $v > x_{ac}$ then
      $\mu_c$ = $\mu_c$ + 1.
    end if
  end for
  $\mu_c$ = $\mu_c$ / $IntervalLength$
  $\mu_a$=$\mu_a$ + $\mu_c$
end for
$\mu_a$ = $\mu_a$ / $AttributeLength$
```

Algorithm 2 Remove percentage x of misleading

```
Input : $Inteval$ values of an attribute $a$ for objects lies in class $c$
Output : $avg$ average of the values of an attribute $a$ in a class $c$
for $x * IntervalLength$ values do
  Remove the value of max difference from the average $avg$ .
end for
```

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. Comparison to other attribute evaluation algorithms

A comparison is applied according to such behavior between different attribute selection algorithms and the proposed interval-based algorithm. The best attribute selection algorithm should follow the following criteria:

- The maximum classification accuracy (peak) is reached with the smallest number of attributes.
- This peak should have the highest value among other attribute evaluation algorithms.

The comparison will be applied through the train and test of the data set multiple times, where in the first time, the data set will contain only the most single relevant attribute, then the number of attributes will be incrementally increasing until the all the attributes are used. This algorithm could be considered as a semi-wrapper method as the evaluation will be applied only on a certain subset of attributes, where the number of this subsets is equal to the number of attributes. The wrapper-based approaches employ induction classifier as a black box using cross-validation or bootstrap techniques. A method has been previously proposed to compare between different attribute selection algorithms through applying genetic algorithm to evaluate the feature subset candidates suggested by different attribute selection algorithms [11]. This algorithm had solved the problem of the high computation but the problem of dealing with the data sets of continuous attributes still a problem in the used attribute selection algorithms.

In the test, two different classifiers will be applied which are the support vector machine (SVM) and multi-layer perceptron. The SVM classifier uses the Gaussian Radial Basis Function kernel as it shows the best classification accuracy. Both have been extensively used as classification tools with a great deal of success from object recognition. The attribute selection algorithms used are Chi-merge, gain ratio and information gain attribute selection algorithms.

Another test is applied on these two data sets used, where the classification test is applied on all the possible combination of attributes. The combination that shows the best classification accuracy is the one generated by the proposed interval-based attribute selection algorithm.

B. Data sets used in classification

The selection of the data sets used are based on the need of a data set of continuous attributes and discrete class label. A medical data sets have used which are considered as real life data sets that have no specific distribution of values and may contain misleading values due to an error in calibrations or collection of data. The first data set used is pima-indians-diabetes data set which is obtained from UCI machine learning repository [12]. The percentage of error in this data set will be considered zero. It consists of 536 objects and 8 attributes. 90% of the input data used for training while the rest of 10% is used in testing. The second data set used is a data about HCV therapy where it is classified according to the response of some patients to the interferon therapy whether they cured or not. It consists of 66 objects and 13 attributes. There is a percentage of error that may occur in this data set, where experts indicate that it falls between 2 to 3%. Due to the low number of objects, 70% of the input data only are used for training while the rest of 30% is used in testing. Both data sets are adjusted such that the number objects in every class is equal in both stages of training and testing, the classification accuracy will be measured by dividing the number of correctly classified objects by the total number of objects in the testing data set.

C. Classification results

1) pima-indians-diabetes data set: When different attribute selection algorithms are used like information Gain (IG), Chi-Merge (CM) and Gain Ratio (GR), these algorithms shows the same order of ranked attributes. This is because these algorithms are all entropy based attribute selection algorithms. The comparison between the order of features ranks of entropy based attribute selection algorithms and the interval-based attribute selection algorithm is shown in table (I). In this table another SVM-based feature selection method (SVMB) [10] is used, where it shows nearly the same results as Information gain algorithm.

TABLE I
THE ORDER OF ATTRIBUTES ACCORDING TO INFORMATION GAIN IG AND INTERVAL-BASED IB FEATURE SELECTION ALGORITHMS

IG	2	8	6	5	1	7	3	4
SVMB	2	6	1	7	8	3	4	5
IB	2	5	7	1	4	6	8	3

Table (II) shows the results when perform classification using Support vector machine (SVM) and Multi-layer perceptron (MLP). The first row in table (II) shows the classification results when using the input data set contains only attribute 2 in the case of using IG and attribute 2 in the case of using IB. The second row the input data set contains attributes 2, 8 in the case of using IG and attributes 2, 5 in the case of IB. The attributes are incrementally increased based on the attribute selection algorithm used until all attributes are used in the input data set.

In the case using MLP classifier, the peak of accuracy has reached with 81.4% when the input data set contain only the first three selected attributes by the IB algorithm which are 2, 5, 7. While the peak when using the other feature selection algorithms is 75.9% and the number of selected attributes is five attributes which are 2, 8, 6, 5, 1. In the case of using SVM both attribute selection algorithms, the proposed IB and the IG attribute selection algorithms, have the same accuracy percentage peak when the first attribute only is selected. It is noticed that both algorithms have selected the same attribute 2 as the most relevant attribute. On the other hand, All the possible combination of attributes are tested in the same way as above using MLP classifier, where 90% of the input data is for training while the rest is for testing. the combination that shows the best results is the set of attributes 2, 5, 7 which is the same set and of the same order generated by the proposed IB algorithm.

TABLE II
CLASSIFICATION RESULTS OF THE PIMA-INDIANS-DIABETES

SVM	SVM	SVM	MLP	MLP	MLP
IG	SVMB	IB	IG	SVMB	IB
64.81	64.81	**64.81**	74.07	74.07	74.07
64.81	61.11	61.11	74.07	74.07	75.92
64.81	**66.66**	62.96	74.07407	70.37	**81.48148**
53.70	66.66	51.85	72.22222	72.22	79.62963
57.40	68.51	57.40	**75.92593**	75.92	75.92593
57.47	62.96	57.40	74.07407	**77.77**	74.07407
55.55	55.55	53.70	72.22222	77.77	68.51852
51.85	51.85	51.85	72.22222	70.37	72.22222

2) HCV data set: Again in the case Information Gain (IG), Chi-Merge (CM) and Gain Ratio (GR) attribute selection algorithms, The attributes are ranked as follows in table (III). The SVM-Based attribute selection shows the same results as the previous algorithms so it is not useful to maintain them in table (III).

TABLE III
THE ORDER OF ATTRIBUTES ACCORDING TO INFORMATION GAIN IG AND INTERVAL-BASED IB FEATURE SELECTION ALGORITHMS

IG	12	13	4	5	1	3	2	9	11	10
IB	3	4	6	13	9	12	1	5	8	7
IG	6	8	7							
IB	11	10	2							

Table (IV) shows the results when perform classification using SVM and MLP after using both entropy based attribute selection algorithm like the IG and the proposed IB algorithms. It shows that in the case of the selected attributes using the proposed IB algorithm, the classification accuracy has the maximum value when using the first four attributes only. Also the peak was 75 % in the case of using MLP, and 65 % in the case of using SVM. So in both classifiers, the selected attributes by IB has a higher peak than those selected by other attribute selection algorithms.

TABLE IV
Classification results of the HCV

SVM	SVM	MLP	MLP
IG	IB	IG	IB
25.0	50.0	37.5	50.0
37.5	37.5	37.5	37.5
37.5	37.5	37.5	37.5
37.5	**75.0**	37.5	**62.5**
62.5	62.5	37.5	50.0
37.5	37.5	37.5	37.5
50.0	37.5	37.5	37.5
50.0	25.0	37.5	37.5
50.0	62.5	37.5	37.5
50.0	62.5	37.5	50.0
50.0	62.5	37.5	37.5
50.0	50.0	37.5	37.5
50.0	50.0	37.5	37.5

IV. Conclusions

The problem of selecting the features that are relevant to the classifier and remove the irrelevant features has been solved using different attribute selection algorithms. The results demonstrate that the proposed interval-based algorithm encouragingly outperforms most of the popular attribute selection algorithms. The proposed algorithm has two phases for feature selection, the first one is to rank all features according the discussed algorithm, then select the subset of features of the highest classification accuracy. the proposed algorithm depends on two conditions to determine the selected subset of attributes, where the conditions are based on how high is the classification accuracy and how less is the number of selected attributes. The algorithm has been applied on a real life data sets, where it leads to the best classification accuracy with the least number of features.

References

[1] C. Shang and Q. Shen,“Aiding classification of gene expression data with feature selection: a comparative study,” Computational Intelligence Research, vol. 1, pp 68–76, 2006.

[2] A. G. K. Janecek, W. N. Gansterer, M. Demel and G. F. Ecker, “On the relationship between feature selection and classification accuracy,” JMLR: Workshop and Conference Proceedings, vol. 4, pp. 90–105, 2008.

[3] Mostafa A. Salama, Aboul Ella Hassanien and Aly A. Fahmy, “Pattern-based Subspace Classification Model,” Second World Congress on Nature and Biologically Inspired Computing (NaBIC), Kitakyushum, Japan, pp. 357–362, Dec. 2010.

[4] M. Prasad, A. Sowmya and I. Koch,“Designing relevant features for continuous data sets using ICA,” Int. J. Comput. Intell. Appl., vol. 7, p.447 , 2008.

[5] Yvan Saeys, Inaki Inza and Pedro Larranaga,“A review of feature selection techniques in bioinformatics,” bioinformatics, Vol. 23, pp. 2507-2517, 2007.

[6] Huan Liu and R. Setiono, “Feature selection via discretization,” IEEE Transactions on Knowledge and Data Engineering, vol. 9, pp. 642–645, Aug. 1997.

[7] Manuel Mejía-Lavalle, Eduardo F. Morales and Gustavo Arroyo,“Two simple and effective feature selection methods for continuous attributes with discrete multi-class,” Lecture Notes in Computer Science, vol. 4827, pp. 452–461, 2007.

[8] W. Duch, T. Wieczorek, J. Biesiada and M. Blachnik, “Comparison of feature ranking methods based on information entropy,” Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 2, pp. 1415-1419, 2004.

[9] Yin-Wen Chang and Chih-Jen Lin, “Feature Ranking Using Linear SVM” , JMLR: Workshop and Conference Proceedings, pp. 53-64, 2008.

[10] J. Weston S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio and V. Vapnik,“Feature Selection for SVMs,” in Proc. Neural Information Processing Systems, pp. 668–674, 2000.

[11] Chi-Ho Tsang, Sam Kwong. and HanliWang,“Genetic-fuzzy rule mining approach and evaluation of feature selection techniques for anomaly intrusion detection,” vol. 40, Issue 9, pp. 2373–2391, Sep. 2007.

[12] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets.html.