A Minority Class Feature Selection Method

German Cuaya, Angélica Muñoz-Meléndez, and Eduardo F. Morales

National Institute of Astrophysics, Optics and Electronics,
Computer Science Department,
Luis Enrique Erro 1, 72840 Tonantzintla, México
{germancs,munoz,emorales}@inaoep.mx
http://ccc.inaoep.mx

Abstract. In many classification problems, and in particular in medical domains, it is common to have an unbalanced class distribution. This pose problems to classifiers as they tend to perform poorly in the minority class which is often the class of interest. One commonly used strategy that to improve the classification performance is to select a subset of relevant features. Feature selection algorithms, however, have not been designed to favour the classification performance of the minority class. In this paper, we present a novel filter feature selection algorithm, called FSMC, for unbalanced data sets. FSMC selects attributes that have minority class distributions significantly different from the majority class distributions. FSMC is fast, simple, selects a small number of features and outperforms in most cases other feature selection algorithms in terms of global accuracy and in terms of performance measures for the minority class such as precision, recall, F-measure and ROC values.

Keywords: feature selection, unbalanced data set, medical domain.

1 Introduction

With the rapid advances in computer and database technologies, data sets with hundreds and thousands of variables or features are now present in pattern recognition, data mining, and machine learning applications [1–4]. Processing such huge data sets is a challenging task because traditional machine learning techniques usually work well only on small data sets. Feature selection addresses this problem by removing irrelevant, redundant, or noisy data. It improves the performance of the learning algorithm, reduces its computational cost and provides better understandings of the produced models [5].

Feature selection algorithms can be widely categorized into two groups: filter and wrapper methods [2, 4, 6–8]. Filter methods evaluate the goodness of the feature subset by using the intrinsic features of the data. They are computationally inexpensive since they do not rely on any induction algorithm. Wrapper methods, on the contrary, directly use the induction algorithm to evaluate the feature subsets. They generally outperform filter methods in terms of prediction accuracy, but are computationally more intensive.

C. San Martin and S.-W. Kim (Eds.): CIARP 2011, LNCS 7042, pp. 417–424, 2011.

The development of our work was motivated by an application in a medical domain with a relatively large number of attributes and a very unbalanced class distribution, that is common to other medical domains, and that poses problems to traditional classification algorithms and to feature selection algorithms that tend not to favour the minority class [9, 10].

There is large number of feature selection algorithms, however, very few research has been targeted particularly towards unbalanced class distributions. In particular [11], the authors propose a performance measure using ROC curves for feature selection. The main disadvantage of this work is that it uses a wrapper approach requiring repetitive and expensive model training during the feature selection process. In [12], the authors modify the ReliefF feature selection algorithm and present three filter-based feature selection techniques to attack unbalanced data sets, namely, give more weight to the instances of the minority class, oversample the minority class or undersample the majority class. The work presented in [13] is more closely related to our work. In that work the authors aproximate the probability density function (PDF) of each feature independently in an unsupervised manner and then removing those features for which their PDFs have higher covering areas with the PDFs of other features which are known as redundant features, it is important to mention that the authors used both majority and minority class data to calculate the PDFs.

In this paper, we propose a novel filter feature selection algorithm named *Feature Selection for Minority Class* (FSMC) that uses the difference between the expected value of the majority class and the expected value of minority class of each attribute to identify the relevant features for the minority class.

We evaluate the efficiency of FSMC by comparing our method to some well-known *filters* and *wrappers* feature selection strategies, applied with five different types of classifiers in several medical data sets from the UCI repository [14] and on a real data set of gait analysis. The results show that FSMC is competitive and in many cases outperforms other features selection algorithms in terms of classification accuracy, precision, recall, F-measure and ROC values for the minority class as well as selected feature size.

The rest of this paper is organized as follows. Section 2 describes the FSMC algorithm. In Section 3 the experimental results are presented, and finally, Section 4 concludes and provides future research directions.

2 FSMC

In this section we introduce a Feature Selection for Minority Class (FSMC) algorithm. The goal of FSMC is to measure the difference between the expected value of the majority class and the expected value of the minority class to select relevant features for classifying the minority class. The rationale behind our proposal is to select those features whose values are particularly different from the values of the majority class and that could help to classify instances from the minority class. The algorithm boils down to obtain the mean and standard deviation of each variable for the majority class and the mean of the same variables for the minority class. If the mean value of the minority class is at least two standard deviations away from the mean value of the majority class, then that feature is selected as relevant. This is a very simple and easy to implement criterion that to our knowledge has not been used before in the literature and, as shown in Section 3, is very competitive with respect to other feature selection algorithms. A description of FSMC is summarized in Algorithm1.

Algorithm 1. The FSMC algorithm

```
begin
Let $Y$ a given set of attributes
Let $Maj(y)$ the majority class data of attribute $y \in Y$
Let $Min(y)$ the minority class data of attribute $y \in Y$
Let $RelAtt = \emptyset$ the output set of relevant attributes calculated by FSMC
for all $y \in Y$ do
    Compute the mean $(\mu_{Maj(y)})$ and standard deviation $(\sigma_{Maj(y)})$ of $y$ in $Maj(y)$
    Compute the mean $(\mu_{Min(y)})$ of $y$ in $Min(y)$
    if $(\mu_{Min(y)} > (\mu_{Maj(y)} + 2 * \sigma_{Maj(y)})) \vee (\mu_{Min(y)} < (\mu_{Maj(y)} - 2 * \sigma_{Maj(y)}))$ then
        Let $RelAtt \leftarrow RelAtt \cup \{y\}$
    end if
end for
Return $RelAtt$
end
```

3 Experimental Results and Discussion

In order to evaluate the performance of our algorithm FSMC, we used five medical data sets from the UCI ML repository, namely, arrhythmia, ozone, Pima Indians diabetes, diabetes and cardio [14]. Additionally, we used information from gait analysis involving elderly subjects provided by researchers of the National Institute of Rehabilitation of Mexico and which motivated the development of this research. In all cases we used a binary class problem.

We used five different classifiers to obtain performance measures over these data sets for the minority class, namely, precision, recall, F-measure and ROC values and also to obtain information from the global accuracy. The selected classifiers were taken from *Weka* [15] and involved different classification strategies with their default parameters: (i) PART (a decision list that uses separate-and-conquer strategy that builds a partial C4.5 decision tree in each iteration and makes the “best” leaf into a rule), (ii) J48 (C4.5 decision tree algorithm), (iii) Bagging (with 10 decision trees classifiers), (iv) BayesLogicRegresion (Bayesian network learning algorithm that estimates the parameters of $P(Y|X)$ using Logistic Regression), and (v) SMO (John Platt’s sequential minimal optimization algorithm for training a support vector classifier).

We compared FSMC against seven feature selection algorithms also taken from Weka[15] with their default parameters, namely, CFsSubsetEval (evaluates a subset of features), FilteredSubsetEval (evaluates a subset of features that has been passed through a filter strategy), SVMattributeEval (evaluates the worth of an attribute using a SVM classifier), Wrapersubseteval (a wrapper feature selection strategy), PrincipalComponents (performs a PCA analysis), InfoGainAttributeEval (uses information gain to select attributes), and ReliefFAttributeEval (implements the ReliefF algorithm).

Table 1 shows in the header row the general characteristics of the different data sets used in these experiments, such as the total number of instances and attributes, as well as the number of instances in the majority and minority classes. This table summarizes also the number of relevant attributes selected by each feature selection algorithms when applied to the different data sets.

Note that FSMC selects fewer relevant attributes than the rest of algorithms in most data sets, with the exception of the human gait data set. This is convenient in problems involving a large number of variables and a few number of instances.

Table 1. Number of variables selected by eight feature selection methods including FSMC

	Datasets					
	Arrhythmia	**Ozone**	**Pima Indians Diabetes**	**Diabetes**	**Gait**	**Cardio**
Instances	273	1876	569	768	270	1831
Attributes	135	72	8	9	31	21
No. Maj. instances	237	1819	500	500	143	1655
No. Min. instances	36	57	69	268	127	176
Feature Selection Algorithms						
CFsSubsetEval	19	18	3	4	2	6
Filteredsubseteval	18	18	3	3	2	3
SVMattributeEval	All	All	All	All	All	All
Wrapersubseteval	None	None	None	None	None	None
PrincipalComponents	50	19	7	All	9	14
InfoGainAttributeEval	All	All	All	All	All	All
ReliefFAttributeEval	All	All	All	All	All	All
FSMC	4	8	1	1	3	1

The global accuracies obtained using the different classifiers in the data sets are shown in Table 2. In this case, we only show the performance of the three best feature selection algorithms. In all the experiments we used 10-fold cross validation.

The results presented in Table 2 show that the classifiers have, in general, better performance with the features selected by FSMC.

Table 3 shows the number of times that the classification of the minority and majority classes of all data sets was better by the different classifiers using the different subsets of attributes. Again the set variables selected by FSMC has in general better performance.

Table 2. Classification accuracy of different data sets with different classifiers based on different set of variables selected by five methods of feature selection including FSMC

	Arrhythmia				
Classifier	All	CFsSubsetEval	Filteredsubseteval	PrincipalComponents	FSMC
PART	89.74	91.58	91.58	86.08	**92.67**
J48	90.84	89.38	89.38	84.25	**92.31**
Bagging	**92.67**	91.94	91.58	90.11	91.94
BayesLogicRegresion	91.58	87.18	87.91	86.81	**92.67**
SMO	86.81	86.81	86.81	86.81	**92.31**
Average	90.33	89.38	89.45	86.81	**92.38**
	Ozone				
PART	95.36	96.54	96.54	96.48	**96.64**
J48	95.63	95.52	95.52	95.95	**96.48**
Bagging	96.86	96.86	96.86	**96.96**	96.80
BayesLogicRegresion	84.22	88.91	88.91	83.69	**90.03**
SMO	**96.96**	**96.96**	**96.96**	**96.96**	**96.96**
Average	93.81	94.96	94.96	94.01	**95.38**
	Gait Analysis				
PART	64.81	77.41	77.41	70.74	**82.22**
J48	69.26	78.15	78.15	78.52	**81.85**
Bagging	69.26	79.63	79.63	74.07	**80.37**
BayesLogicRegresion	**65.93**	53.33	53.33	57.41	59.63
SMO	52.22	**63.70**	**63.70**	54.44	54.44
Average	64.30	70.44	70.44	67.04	**71.70**
	Pima Indians Diabetes				
PART	87.70	88.23	88.23	89.10	**89.63**
J48	88.40	88.93	88.93	88.05	**89.63**
Bagging	87.17	88.23	88.23	88.40	**88.93**
BayesLogicRegresion	87.87	87.87	87.87	87.87	87.87
SMO	87.87	87.87	87.87	87.87	**88.75**
Average	87.80	88.22	88.22	88.26	**88.96**
	Diabetes				
PART	73.05	72.27	**73.31**	73.05	72.01
J48	71.48	73.44	**75.26**	71.48	72.01
Bagging	**76.69**	75.52	75.00	**76.69**	71.88
BayesLogicRegresion	**65.76**	63.93	64.06	**65.76**	65.10
SMO	65.10	62.89	63.54	65.10	**69.27**
Average	**70.42**	69.61	70.23	70.42	70.05
	Cardio				
PART	**98.74**	98.31	97.00	98.69	93.99
J48	98.53	**98.74**	97.27	98.03	93.99
Bagging	98.47	**98.53**	97.21	98.03	93.99
BayesLogicRegresion	93.56	91.43	93.17	91.86	**93.99**
SMO	91.59	92.41	**95.79**	93.66	93.99
Average	**96.18**	95.88	96.09	96.06	93.99

Table 3. Times that the classification of majority and minority class was better using different set of variables

	ALL	CfsSubsetEval	Filteredsubseteval	PrincipalComponents	FSMC
Wins in Accur. Min. class	8	12	8	6	**13**
Wins in Accur. Maj. class	6	8	6	11	**20**

Table 4. Times that the Precision, Recall, F-Measure, and ROC measures are better for the different feature selection algorithms

		ALL			CfsSubsetEval			Filteredsubseteval			PrincipalComponents			SFMC		
	Measure	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
Arrhythmia	Precision	2	0	3	0	0	5	0	0	5	0	0	5	3	0	2
	Recall	1	1	3	0	2	3	0	1	4	0	0	5	2	1	2
	F-Measure	2	0	3	0	1	4	0	1	4	0	0	5	2	0	3
	ROC	1	0	4	1	1	3	0	1	4	0	0	5	2	0	3
	Sum	6	1	13	1	4	15	0	3	17	0	0	20	**9**	1	10
	Perc.	30.00%	5.00%	65.00%	5.00%	20.00%	75.00%	0.00%	15.00%	85.00%	0.00%	0.00%	100.00%	**45.00%**	5.00%	50.00%
Ozone	Precision	0	2	3	0	5	0	0	5	0	0	2	3	0	2	3
	Recall	1	2	2	0	4	1	0	4	1	0	2	3	0	2	3
	F-Measure	1	2	2	0	4	1	0	4	1	0	2	3	0	2	3
	ROC	0	1	4	0	3	2	0	3	2	1	1	3	1	1	3
	Sum	**2**	7	11	0	16	4	0	16	4	1	7	12	1	7	12
	Perc.	**10.00%**	35.00%	55.00%	0.00%	80.00%	20.00%	0.00%	80.00%	20.00%	5.00%	35.00%	60.00%	5.00%	35.00%	60.00%
Gait	Precision	0	0	5	0	2	3	0	2	3	0	0	5	3	0	2
	Recall	1	0	4	0	1	4	0	1	4	0	0	5	3	0	2
	F-Measure	1	0	4	0	1	4	0	1	4	0	0	5	3	0	2
	ROC	1	0	4	0	2	3	0	2	3	0	0	5	2	0	3
	Sum	3	0	17	0	6	14	0	6	14	0	0	20	**11**	0	9
	Perc.	15.00%	0.00%	85.00%	0.00%	30.00%	70.00%	0.00%	30.00%	70.00%	0.00%	0.00%	100.00%	**55.00%**	0.00%	45.00%
Pima Indians Diabetes	Precision	0	1	4	0	1	4	0	1	4	0	1	4	4	1	0
	Recall	0	1	4	0	2	3	0	2	3	1	1	3	2	1	2
	F-Measure	0	1	4	0	2	3	0	2	3	1	1	3	2	1	2
	ROC	2	1	2	0	1	4	0	1	4	0	1	4	2	1	2
	Sum	2	4	14	0	6	14	0	6	14	2	4	14	**10**	4	6
	Perc.	10.00%	20.00%	70.00%	0.00%	30.00%	70.00%	0.00%	30.00%	70.00%	10.00%	20.00%	70.00%	**50.00%**	20.00%	30.00%
Diabetes	Precision	0	2	3	0	0	5	0	1	4	0	2	3	2	1	2
	Recall	0	2	3	2	0	3	0	0	5	0	2	3	1	0	4
	F-Measure	0	2	3	1	0	4	1	0	4	0	2	3	1	0	4
	ROC	0	3	2	1	2	2	0	0	5	0	3	2	1	0	4
	Sum	0	9	11	4	2	14	1	1	18	0	9	11	**5**	1	14
	Perc.	0.00%	45.00%	55.00%	20.00%	10.00%	70.00%	5.00%	5.00%	90.00%	0.00%	45.00%	55.00%	**25.00%**	5.00%	70.00%
Cardio	Precision	0	1	4	0	2	3	0	0	5	0	1	4	3	2	0
	Recall	0	0	5	2	0	3	1	0	4	1	0	4	1	0	4
	F-Measure	0	0	5	2	0	3	1	0	4	1	0	4	1	0	4
	ROC	0	0	5	2	0	3	1	0	4	1	0	4	1	0	4
	Sum	0	1	19	6	2	12	3	0	17	3	1	16	**6**	2	12
	Perc.	0.00%	5.00%	95.00%	30.00%	10.00%	60.00%	15.00%	0.00%	85.00%	15.00%	5.00%	80.00%	**30.00%**	10.00%	60.00%

Table 5. Summary of winners of Table 4

	ALL	CfsSubsetEval	Filteredsubseteval	PrincipalComponents	FSMC
Precision	2	0	0	0	**15**
Recall	3	4	1	2	**9**
F-Measure	4	3	2	2	**9**
ROC	4	4	1	2	**9**

Finally, Table 4 shows complementary information about the effectiveness of FSMC on the minority class on the six data sets. This table show how many times the precision, recall, F-measure and ROC values were better on these measures for the minority class with the classifiers used with specific set of variables. Table 5 shows the summary of how many times each feature selection algorithm won over the other algorithms in the different performance measures shown in Table 4. Again, FSMC outperforms the other feature selection algorithms in all of these measures.

4 Conclusions and Future Work

In this paper, we have presented a novel feature selection algorithm useful for unbalanced data sets. Its main feature selection strategy is based on selecting those features whose values are particularly different from the values of the majority class and that could help to classify instances from the minority class.

The experimental results show that the proposed method tends to select fewer attributes than other feature selection methods and, at the same time, outperforms most of the time such algorithms in different performance measures when tested on several data sets and with different classification algorithms.

As part of the future work we would like to extend the selection strategy to nominal attributes. We would also like to extend the selection strategy to real-valued data that do not follow a Gaussian distribution.

References

1. Jain, A., Zongker, D.: Feature Selection: Evaluation, Application, and Small Sample Performance. IEEE Trans. Pattern Analysis and Machine Intelligence 19(2), 153–158 (1997)
2. Dash, M., Liu, H.: Feature Selection for Classification. Intelligent Data Analysis 1(3), 131–156 (1997)
3. Dash, M., Liu, H.: Consistency-based Search in Feature Selection. Artificial Intelligence 151(1-2), 155–176 (2003)
4. Kohavi, R., John, G.H.: Wrapper for Feature Subset Selection. Artificial Intelligence 97(1-2), 273–324 (1997)
5. Liu, H., Motoda, H.: Feature Selection for Knowledge Discovery and Data Mining. Kluwer Academic Publishers, Norwell (1998)
6. Robnic-Sikonja, M., Kononenko, I.: Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning 53(1-2), 23–69 (2003)

7. Mao, K.Z.: Feature Subset Selection for Support Vector Machines Through Discriminative Function Pruning Analysis. IEEE Transactions on System, Man and Cybernetics, Part B 34(1), 60–67 (2004)
8. Hsu, C.N., Huang, H.J., Dietrich, S.: The ANNIGMA-Wrapper Approach to Fast Feature Selection for Neural Nets. IEEE Transactions on System, Man and Cybernetics, Part B 32(2), 207–212 (2004)
9. Japkowicz, N., Stephen, S.: The Class Imbalance Problem: A Systematic Study. Intelligent Data Analysis 6(5), 429–449 (2002)
10. Weiss, G.M., Provost, F.: The effect of class distribution on classifier learning: an empirical study. Technical report, Department of Computer Science, Rutgers University, New Jersey (2001)
11. Chen, X., Wasikowski, M.: FAST: A ROC-based feature selection metric for small samples and imbalanced data classification problems. In: 14th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 124–132 (2008)
12. Kamal, A.H.M., Zhu, X., Pandya, A.S., Hsu, S., Narayanan, R.: Feature Selection for Datasets with Imbalanced Class Distributions. International Journal of Software Engineering and Knowledge Engineering 20(2), 113–137 (2010)
13. Alibeigi, M., Hashemi, S., Hamzeh, A.: Unsupervised Feature Selection Based on the Distribution of Features Attributed to Imbalanced Data Sets. International Journal of Artificial Intelligence and Expert Systems 2(1), 133–144 (2011)
14. Frank, A., Asuncion, A.: UCI Machine Learning Repository. University of California, School of Information and Computer Science, Irvine, CA (2010), `http://archive.ics.uci.edu/ml`
15. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)