Vol.1 pp. 20-25 **Journal of Operational Management & Marketing Strategies (OMMS)**

Rank Aggregation for Filter Feature Selection in Credit Scoring

Bouaguel Waad
LARODEC
ISGT, University of Tunis
Email: bouaguelwaad@mailpost.tn

Bel Mufti Ghazi
LARIME
ESSEC, University of Tunis
Email: belmufti@yahoo.com

Limam Mohamed
LARODEC
ISGT, University of Tunis, Dhofar University, Oman
Email: mohamed.limam@isg.rnu.tn

***Abstract*—The credit industry is a fast growing field, credit institutions collect data about credit customer and use them to build credit model. The collected information may be full of unwanted and redundant features which may speed down the learning process, so, effective feature selection methods are needed for credit dataset. In general, Filter feature selection methods outperform other feature selection techniques because they are effective and computationally fast. Choosing the appropriate filtering method from the wide variety of classical filtering methods proposed in the literature is a crucial issue in machine learning. So, we propose a feature selection fusion model that fuses the results obtained by different filter feature selection methods via aggregation techniques. Evaluations on four credit datasets show that the fusion model achieves good results.**

I. Introduction

Many empirical studies show that manipulating few variables in credit scoring leads certainly to more reliable and better understandable models without irrelevant, redundant and noisy data [1]. The more the number of features grows the more computation is required and model accuracy and scoring interpretation are reduced [2]. To overcome these problems we perform a feature selection on the original features set.

In feature selection process we choose an appropriate feature subset that contains the most relevant features. A variety of techniques to select the best subset of features have been proposed. Three main classes of feature selection are identified in the literature as stated by [3], [4]: filter, wrapper and hybrid feature selection methods. A filter technique is a pre-selection process which is independent of the later applied classification algorithm. Filters can be exceptionally effective because they need to be performed only once without any search involved. A wrapper technique on the other hand uses specific classifier and exploits resulting classification performance to select features. This kind of methods use search techniques to pick subsets of variables and evaluate their importance based on the estimated classification accuracy [4]. The hybrid approach uses both filtering and wrapping methods for improving the performance of the feature selection.

According to [5] filters methods outperforms other feature selection methods in many cases. There are a variety of classical filter methods in previous literature [1], [6]. Given the variety of techniques, the question is how to choose the best one for a specific feature selection task? [5] call this problem a selection trouble. Hence, we propose to investigate on a new fusion framework. In this paper we focus on combining different filtering criteria into a new result in order to obtain a better rank list, by using a aggregation rules. This paper is organized as follows. Section 2 reviews filter feature selection methods and features aggregation. Section 3 gives experimental results on four datasets and in Section 4 conclusions are drawn.

II. Selection Trouble

A. Filter Feature Selection Method

The basic idea of filter methods is to select the best features according to some prior knowledge. Filter feature selection methods can be grouped into two categories, i.e. feature weighting methods and subset search methods. This categorization is based on whether they evaluate the relevance of features separately or through feature subsets. In feature weighting methods, weights are assigned to each feature independently and then the features are ranked based on their relevance to the target variable. Relief is a famous algorithm that study features relevance [7]. This method uses the Euclidean distance to select a sample composed of a random instance and the two nearest instances of the same and opposite classes. Then a routine is used to update the feature weight vector for every sample triplet and determines the average feature weight vector relevance. Then, features with average weights over a given threshold are selected.
Subset search methods explore all possible feature subsets using a particular evaluation measure. The best possible subset is selected when the search stops. According to [8], consistency and correlation [9], [10] are the best evaluation measures that decrease efficiently irrelevance and redundancy. A Consistency measure evaluates the distance of a feature subset from the consistent class label. Consistency is established when a data set with the selected features alone is consistent. That is, no two instances may have the same feature values if they have a different class label [10]. A correlation measure is applied between two features as a goodness measure. That is a feature is considered as good if it is highly correlated to the class and uncorrelated with any other features. [8] recommended two main approaches to measure correlation, the first one is based on classical linear correlation between to random variables and the second one is based on information theory.

ISSN: 2961-662X

Numerous correlation coefficients can be used under to first approach but the most common is the Pearson correlation coefficient (PCC). PCC is a simple measure that has been shown to be effective in a wide variety of feature selection methods ([4]). Formally, the PCC for two continuous random variables x_i and x_j is defined as :

$$PCC = \frac{cov(X_i, X_j)}{var(X_i)var(X_j)}, \quad (1)$$

where where cov is the covariance of variables and var is the variance of each variable. Simple correlation measure in general measures the linear relationship between two random variables, which may be not suitable in some cases. The second approach based on information theory measures how much knowledge two variables carry about each other. Mutual information (MI) is a well known information theory measure that captures nonlinear dependencies between variables. Formally, the mutual information of two continuous random variables x_i and x_j is defined as:

$$I(x_i, x_j) = \int \int p(x_i, x_j) log \frac{p(x_i, x_j)}{p(x_i)p(x_j)} dx\, dx, \quad (2)$$

where $p(x_i, x_j)$ is the joint probability density function, and $p(x_i)$ and $p(x_j)$ are the marginal probability density functions.

The majority of above cited features selection methods select the k top ranked features. In general, filter criteria are used independently. That is, one feature selection method is employed and performance is measured according to the selected features. The question is then which method will be the most appropriate to our study. Rather than to study what each single criterion can offer, we can employ these methods in combination.

III. Ensemble Features Methods

A. Filter Feature Selection Aggregation

Two effective modes to fuse a set of filtering feature selection methods are proposed in the literature [5]. In the first mode, the final outputs of each single filter method are combined into a one single result. The second fusion mode, combine the different filtering criteria of each filter method in order to find a new measure that select the best feature subset. In general, when the second mode is used, we not only need some prior knowledge about the data but also a familiarity with the criteria to be combined and good mathematical skills, therefore the first mode is choose over the second, because it is the simplest one and because it does not require additional configuration. In order to implement the chosen fusion mode, aggregation techniques can be used.

The main thought behind using ensemble feature aggregation is to obtain a list of significant and jointly selected set of features that can be used during the classification process. We try in this context to capture features which may provide essential factors during the prediction of the credit-worthiness by removing the redundant ones. Typically ensemble feature aggregation reduce the biases caused by individual feature algorithms while providing higher accuracy, sensitivity, and specificity, which are often not achievable with individual feature selection techniques or while not using any feature selection techniques at all.

In general, when we deal with aggregating feature rankings, there are two issues to consider. The first one is which base feature rankings to aggregate. There are different ways to generate the base feature rankings:

- using the same dataset but by different filter methods.
- using different datasets but the same filtering method.
- using different subsamples of the same dataset and the same ranking method.

The second issue concerns the type of aggregation function to use. Ensemble selection consists of multiple runs of feature ranking which are then combined into a single ranking for each feature. One of the most critical decisions when performing ensemble feature selection is deciding on which aggregation technique to use for combining the resulting ranked feature lists from the multiple runs of feature ranking into a single decision for each feature.

For the first issue we decide to use the same dataset with different filter methods. The three previously discussed feature selection criteria namely relief, PCC and MI are then considered. For the second issue many functions are available in the literature, like taking the mean or median of the ranks. This paper is an in-depth comparison between two aggregation techniques: Majority Vote and Mean Aggregation.

Majority vote is a common classifier combination method, particularly used in classifier ensembles when the class labels of the classifiers are crisp [11]. In general, majority voting is a simple method that does not require any parameters to be trained or any additional information for the later results [3]. We propose to use majority voting to feature selection in order to fuse an ensemble of filter methods. This method use voting for selecting the features with the major amount of votes. In this case the input is a set of ranking lists generated by several feature selection techniques, and which are sorted in descending order according to their corresponding votes, from the most significant feature to the least one. The output is a single list of features corresponding to the most discriminating features.

Mean Aggregation consists of taking the average rank across all of the ranked feature lists and using that mean value to determine the final rank of the feature. Mean aggregation technique is practical and easy to implement which make it frequently used for ensemble feature selection [12].

B. Error Curve

Once the selection trouble is resolved and a consensus list of mutual features is obtained, we come across the issue of choosing the appropriate number of features to retain. In fact a list of sorted features doesn't provide us with the optimal features subset. In general a predefined small number of features is retained from the consensus list for constructing

ISSN: 2961-662X

the final model. If the number of used features is relatively small or big, then the final classification results may be a degraded

In this section, we approach the problem of choosing the appropriate number of features by following the idea that the precision of the feature rank is related to predictive accuracy. In fact aggregation would put on top of a list a feature that is most important, and at the bottom a feature that is least important relatively to the target concept. All the other features would be in-between, ordered by decreasing importance. By following this intuition, we choose the number of the most pertinent features by performing a stepwise feature subset evaluation, with which we generate a so-called error curve. We rely on the process of generating the error curve (Figure 1). We begin with the obtained ranked list in Section III, we then construct the credit model with only the top-ranked feature and we then add to this feature the second ranked feature . This process is continued iteratively until a bottom ranked feature is added yielding to decrease in the general accuracy. The points of the error curve are each of the n estimated errors and the point where the error curve decrease is considered as the selection boundary for the appropriate number of features.

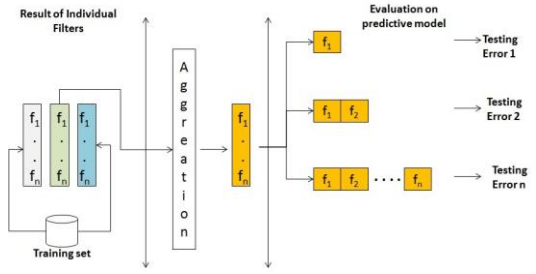

Fig. 1. Ensemble feature selection

IV. Empirical Study

four real-world datasets with detailed input attributes description are selected to study the performance of the proposed approach: two datasets from the UCI repository of machine learning databases (i.e. Australian and German credit datasets) and a dataset from a Tunisian bank and the HMEQ dataset.

- The Australian dataset present an interesting mixture of attributes: 6 continuous, 7 nominal and a target attribute with few missing values. This dataset is composed of 690 instances where 306 ones are creditworthy and 383 are not. All attribute names and values have been changed to meaningless symbols for confidentiality.
- The German credit dataset covers a sample of 1000 credit consumers where 700 instances are creditworthy and 300 are not. For each applicant, 21 numeric input variables are available .i.e. 7 numerical, 13 categorical and a target attribute.
- The HMEQ dataset covers a sample of 5960 instances describing recent home equity loans where 4771 instances are creditworthy and 1189 are not. The target is a binary variable that indicates if an applicant is eventually defaulted. For each applicant, 12 input variables were recorded where 10 are continuous features, 1 is binary and 1 is nominal.
- The Tunisian dataset covers a sample of 2970 instances of credit consumers where 2523 instances are creditworthy and 446 are not. Each credit applicant is described by a binary target variable and a set of 22 input variables were 11 features are numerical and 11 are categorical. Table I displays the characteristics of the datasets that have been used for evaluation.

TABLE I
Results summary for the Australian dataset.

	TP Rate	Precision	Recall	F-Measure
	LR			
Relief	0.923	0.923	0.923	0.923
PCC	0.924	0.926	0.924	0.926
MI	0.944	0.919	0.944	0.929
Majority	0.946	0.926	0.946	0.934
Mean	0.934	0.927	0.934	0.931
	NB			
Relief	0.88	0.941	0.88	0.909
PCC	0.935	0.918	0.935	0.927
MI	0.944	0.903	0.944	0.923
Majority	0.945	0.948	0.932	0.929
Mean	0.943	0.923	0.943	0.928
	SVM			
Relief	0.880	0.941	0.88	0.909
PCC	0.880	0.931	0.86	0.905
MI	0.890	0.910	0.908	0.890
Majority	0.908	0.931	0.908	0.910
Mean	0.908	0.931	0.890	0.910

In general mutual information computation requires estimating density functions for continuous variables. For simplicity, each variable is discretized. Then, we split the datasets into a training sample and a test sample, where the first deals with the new feature selection approach and the diverse classification models and the second one checks the reliability of the constructed models in the learning step. The experimental study compares the performance of the fusion approach with the individual filter methods. The performance of our system is evaluated using the True positive (TP) and False positive (FP) rates and the standard Information retrieval (IR) performance measures: Precision, Recall and F-measure metrics. Results summarized in each Table I and Table II represent the performance of each feature selection technique for three different classification techniques: Logistic Regression (LR), Naive Bayes (NB) and Support Vector Machine (SVM).

Tables I-IV summarize the performances achieved by LR, NB, and SVM algorithms using 3 individual filters namely relief, PCC, MI and their majority vote aggregation and mean aggregation. A more detailed picture of the achieved results shows that in most cases, aggregation approaches usually outperform single filters.

ISSN: 2961-662X

TABLE II
RESULTS SUMMARY FOR THE GERMAN DATASET.

	TP Rate	Precision	Recall	F-Measure
	LR			
Relief	0.511	0.692	0.511	0.588
PCC	0.5	0.721	0.500	0.591
MI	0.580	0.750	0.580	0.654
Majority	0.578	0.781	0.586	0.658
Mean	0.578	0.781	0.586	0.656
	NB			
Relief	0.5	0.638	0.5	0.561
PCC	0.477	0.737	0.477	0.579
MI	0.523	0.742	0.523	0.613
Majority	0.556	0.716	0.545	0.619
Mean	0.542	0.750	0.542	0.612
	SVM			
Relief	0.489	0.694	0.489	0.573
PCC	0.489	0.705	0.489	0.577
MI	0.545	0.738	0.545	0.627
Majority	0.557	0.766	0.557	0.645
Mean	0.552	0.766	0.552	0.627

TABLE III
RESULTS SUMMARY FOR THE HMEQ DATASET.

	TP Rate	Precision	Recall	F-Measure
	LR			
Relief	0.836	0.819	0.836	0.81
PCC	0.974	0.838	0.974	0.901
MI	0.836	0.819	0.836	0.81
Majority	0.968	0.853	0.976	0.912
Mean	0.966	0.850	0.966	0.904
	NB			
Relief	0.8	0.747	0.8	0.736
PCC	0.832	0.818	0.832	0.798
MI	0.831	0.814	0.831	0.801
Majority	0.97	0.843	0.97	0.902
Mean	0.981	0.821	0.981	0.887
	SVM			
Relief	0.807	0.845	0.807	0.728
PCC	0.828	0.822	0.828	0.784
MI	0.828	0.822	0.828	0.784
Majority	0.989	0.835	0.989	0.905
Mean	0.987	0.830	0.987	0.902

V. CONCLUSION

In this study, we investigate on merging filter feature selection methods within a credit scoring framework. Our work was conducted on two parts. First, we conducted a preliminary study on two rank aggregation approaches, namely majority voting and mean aggregation. Second we investigated on choosing the right number of features from the final ranked list, we evaluated the ranking by performing a stepwise feature subset evaluation, resulting on an error curve. Results show that there is a generally beneficial effect of aggregating feature rankings as compared to the ones produced by single methods.

TABLE IV
RESULTS SUMMARY FOR THE TUNISIAN DATASET.

	TP Rate	Precision	Recall	F-Measure
	LR			
Relief	0.848	0.827	0.847	0.830
PCC	0.850	0.833	0.850	0.832
MI	0.852	0.822	0.852	0.826
Majority	0.985	0.866	0.985	0.921
Mean	0.964	0.875	0.964	0.917
	NB			
Relief	0.888	0.876	0.888	0.882
PCC	0.880	0.876	0.880	0.879
MI	0.883	0.885	0.883	0.884
Majority	0.981	0.866	0.981	0.920
Mean	0.960	0.860	0.962	0.913
	SVM			
Relief	0.85	0.722	0.85	0.781
PCC	0.847	0.769	0.847	0.784
MI	0.994	0.851	0.994	0.917
Majority	0.998	0.849	0.999	0.930
Mean	0.993	0.840	0.994	0.927

In fact the fusion performance is either superior to or at least as close as either of filter methods. In additional to this work, selecting the right number of features is a challenge, however to select the appropriate number of feature from a ranking list is still an open problem to be studied in the future. In our further work we plan to go beyond the visual inspection of the error curves. The first step would be to use the area under the error curve as a metric to evaluate the quality of the curves.

REFERENCES

[1] C. M. Wang and W. F. Huang, "Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data," *Expert Syst. Appl.*, vol. 36, pp. 5900–5908, 2009.

[2] T. Howley, M. G. Madden, M. L. O'Connell, and A. G. Ryder, "The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data." *Knowl.-Based Syst.*, vol. 19, pp. 363–370, 2006.

[3] E. Guldogan and M. Gabbouj, "Feature selection for content-based image retrieval," *Signal, Image and Video Processing*, pp. 241–250, 2008.

[4] I. Rodriguez, R. Huerta, C. Elkan, and C. S. Cruz, "Quadratic Programming Feature Selection," *Journal of Machine Learning Research*, vol. 11, pp. 1491–1516, Apr. 2010.

[5] O. Wu, H. Zuo, M. Zhu, W. Hu, J. Gao, and H. Wang, "Rank aggregation based text feature selection," in *Web Intelligence*, 2009, pp. 165–172.

[6] W. Bouaguel and G. Bel Mufti, "An improvement direction for filter selection techniques using information theory measures and quadratic optimization," *International Journal of Advanced Research in Artificial Intelligence*, vol. 1, pp. 7–11, August 2012.

[7] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the ninth international workshop on Machine learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992, pp. 249–256.

[8] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *ICML*, 2003, pp. 856–863.

[9] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *ICML*, 2000, pp. 359–366.

[10] A. Arauzo-Azofra, J. M. Benitez, and J. L. Castro, "Consistency measures for feature selection," *J. Intell. Inf. Syst.*, vol. 30, no. 3, pp. 273–292, Jun. 2008.

ISSN: 2961-662X

[11] L. I. Kuncheva, J. C. Bezdek, and P. W. Duin, "Decision templates for multiple classifier fusion: an experimental comparison," *Pattern Recognition*, vol. 34, pp. 299–314, 2001.
[12] R. Wald, T. M. Khoshgoftaar, and D. J. Dittman, "Mean aggregation versus robust rank aggregation for ensemble gene selection," in *ICMLA (1)*, 2012, pp. 63–69.

ISSN: 2961-662X