Improving EM clustering algorithm by using stochastic local search method: application to credit risk assessment in banking and finance

Abdullah A.K. Alkhawaldeh[1] and Dalila Boughaci[2]
{alkwaldhh@yahoo.com, dboughaci@usthb.dz}

[1]Department of Accounting, Faculty of Economics and Administrative Sciences, The Hashemite University, Zarqa, Jordan. [2]LRIA-FEI- USTHB- Computer Science Department, BP 32 El-Alia Bab-Ezzouar, 16111, Algiers, Algeria

Abstract. Data mining is the process of analyzing large datasets in order to discover patterns and extract useful information. Clustering and classification are the most prominent tasks of data mining. Classification is a supervised learning algorithm that can be used to classify data with known class labels. Clustering is an unsupervised learning algorithm without predefined class labels. Clustering is used to split a large dataset into groups or clusters. In this paper, we propose a novel approach that uses a stochastic local search based feature selection method to improve the Expectation-Maximization (EM) clustering algorithm. The aim is to identify an optimum feature subset and to increase the accuracy rate in credit risk assessment. This technique may be used to help banks and managers in decision-making. The proposed method is evaluated on several financial datasets. The numerical results are promising and show the benefits of the proposed method in clustering the considered data for credit risk assessment.

Keywords: Clustering, feature selection, stochastic local search, credit risk assessment.

1 Introduction

Credit risk management is the process that examines the practices and techniques that may be used to manage financial risks. To assess credit risk, several models have been used such as judgmental methods and credit risk rating based models. The judgmental methods are based mainly on assessor's experience to make a decision whether to accept or refuse credit. Credit rating based models use generally statistical methods to derive an appropriate decision [1, 12]. Credit rating has several benefits on banks and financial institutions. It helps in managing loan portfolios and in reducing the cost of the credit evaluation which may enhance the credit decision-making. The credit evaluation can be defined as the process that evaluates the loans creditworthy based on some related variables and decides whether accepting or rejecting the loan's credit [11]. Several techniques have been developed for building credit scoring models. Among them, we cite: the statistical techniques such as Support vector machines [3], Classification and Regression Trees [7] Bayesian networks [10], Neural networks [9], Random forest [13], Ensemble classifiers [2] and Local search methods [5, 6].

In this paper, we use clustering data mining technique to assess credit risk. We propose a new technique to cluster and classify creditworthy loans against non-credit worthy ones which can reduce the lender's credit risk. A cluster is a group of similar patterns sharing a

InCOGITE 2019, November 07-09, Tangerang, Indonesia

DOI 10.4108/eai.7-11-2019.2295238

same characteristic. We evaluate the Expectation-Maximization (EM) algorithm [8] on several datasets. Then, we propose a stochastic local search based feature selection to increase the performance of EM and improve its efficiency.

EM is a hierarchical clustering algorithm able to produce informative clusters consisting of similar patterns also called data points. However, EM is not suitable for large datasets with a large number of features. In this work, we propose an improved EM clustering algorithm by using a feature selection method. The proposed feature selection is based on the stochastic local search. The latter is an iterative method able to select a good subset of features for data analysis. Feature selection is an important and a useful pre-processing step for data analysis in particular for datasets with large numbers of features. The aim is to select the best features to form good quality clusters having a high degree of similarity and improve the clustering performance. The proposed method is evaluated on some well-known credit datasets. When EM is combined with the stochastic local search based feature selection, the likelihood of the obtained model of the data is maximized. This indicates an increase in the degree of similarity of the points within the same cluster. That means that the points, in each cluster, are similar as possible, and those in different clusters are as different as possible. Also, the accuracy rate is improved. The rest of this paper is organized as follows: Section-2 gives a background of the main concepts used in this study. Section-3 details the proposed approach. Section-4 presents the empirical studies. Finally Section-5 concludes and gives some future works.

2 Background

The aim of this section is to give a background of the main concepts used in our study.

2.1 Classification and clsutering

Data mining is the process that permits to discover patterns in large datasets by using a set of techniques and algorithms. Clustering and classification are the most important means of data mining. Classification is a supervised learning technique used to classify data with known class labels. Clustering is an unsupervised learning algorithm that may be used to split a large dataset into groups also called clusters. The aim of clustering is to find homogeneous groups such that objects in the same group are more similar to each other than the other. The clustering technique is an interesting tool that can help banks and managers in decision-making. The process is to partition a large dataset (in our case a set of loans) into groups. It can help to distinguish between good and bad loans in terms of their creditworthiness. Each loan is evaluated according to its profile. For instance, a profile of a given client is computed by using a set of descriptive variables or features such as: the age of client, his salary, his historical payments, the guarantees, and default rates.

In this work, we are interested in the Expectation-Maximization (EM) algorithm [8]. The EM clustering is an iterative algorithm that operates on two main steps: Expectation and Maximization. In the expectation step, we compute the cluster probabilities. The maximization step aims to maximize the likelihood of the model of the data. In our case, we have two clusters to be created. EM starts with a random model. It computes, for each instance, the probability of belonging to each of the two clusters. The probabilistic description of the clusters is computed based on mean and standard deviation for the numerical variables. This process of probabilities computation is repeated until convergence.

2.2 Feature Selection

Feature selection is a pre-processing step that can be used with data mining techniques to enhance their performance. It helps in removing the redundant features deemed irrelevant to the data mining task. Feature selection can reduce the data dimensionality and improve data mining efficiency. This can be done by identifying an optimum feature subset increasing the accuracy rate. The identified features are used to build a model for data classification or clustering. There are two main feature selection methods which are: the wrapper methods and the filtering methods [6]. The filtering methods are based on heuristics. They eliminate and filter out the undesirable features before launching the data mining task. The filtering methods use statistical feature of the data as an evaluation measure rather than using a learning algorithm used with the wrapper methods. The wrapper methods are based generally on machine learning algorithm for searching the best subset of features. The machine learning algorithm selects the best set of features with high accuracy from the obtained confusion matrix. This make the wrapper methods computationally expensive compared to the filtering methods.

2.3 Stochastic local search

The stochastic local search (SLS) is an iterative algorithm that has been used successfully for solving various optimization problems [4]. The algorithm starts with an initial random solution. Then, it explores the search space looking for new good solutions. The method combines both diversification and intensification strategies to locate good solutions in the search space. The diversification strategy consists in generating random neighbor solutions. The intensification strategy consists in finding best neighbor solutions according to an objective function or an evaluation measure. The intensification phase is applied with a fixed probability $wp > 0$ and the diversification phase is applied with a probability $1 - wp$. The wp is a probability fixed empirically. The iterative process of SLS is repeated until a certain number of iterations or a criterion is reached.

3. Proposed Approach

We propose a stochastic local search (SLS) based feature selection method to find the optimal subsets of features to be used in the clustering step. The role of SLS is to find optimal combinations of features from the dataset. The potential features are used with the EM clustering where the aim is to obtain good clusters and to improve the clustering efficiency. SLS is used for feature selection in order to remove the redundant features deemed irrelevant to the data clustering task. The different components of our approach are detailed in the next subsections.

4.1 Solution representation

Feature selection can be viewed as an optimization problem where a solution is a set of potential features increasing the accuracy rate in credit risk assessment and the objective function value is the accuracy rate. To represent a possible solution of this problem, we use a binary vector with the length of the vector equal to n, where n is the number of features. When a feature is selected, the value 1 is assigned to it; a value 0 is assigned to it otherwise. Figure-

1 depicts an example of vector for a data with eleven features where the first, the second, the third, the sixth, the ninth, the tenth and the eleventh features are selected. The others are not.

1	1	1	0	0	1	0	0	1	1	1

Figure 1. Example of a Solution representation

4.2 SLS for feature selection

The SLS feature selection method starts with an initial solution considering all the features. Then it applies both diversification and intensification strategies to generate new solutions. With the diversification strategy, the neighbors' solutions are generated by randomly adding or deleting a feature. With the intensification strategy, the neighbor solution x' of a current solution x is obtained by modifying one bit. For a candidate solution, the method generates all the neighbor solutions of this candidate solution. Then it selects the best one to be the next candidate solution for the next iteration. The quality of a solution is measured by using the accuracy rate. The intensification step is applied with a fixed probability $wp > 0$ and the diversification step with a probability (1-*wp*). The *wp* is a probability fixed empirically. The SLS feature selection is combined with the EM algorithm for clustering data. The overall method combined both SLS and EM algorithm, is repeated for a certain number of iterations (*max_iterations*) fixed empirically. The EM clustering uses the Log-likelihood as a measure of variation within a cluster. The Log-likelihood is used to estimate the distribution of data points. The EM algorithm aims to maximize the probability (likelihood) of every data point. For SLS, the solution quality is measured by using an objective function given as:

f(x) = Accuracy= (TP+TN)/(TP+TN+FP+FN) where True Positives (TP): is the number of positive examples, labeled as such. False Positives (FP): is the number of negative examples, labeled as positive. True Negatives (TN): is the number of negative examples, labeled as such. False Negatives (FN): is the number of positive examples, labeled as negative.

4.3 The overall algorithm

The overall method combined SLS with EM algorithm is sketched in Algorithm 1.

Algorithm 1. SLS with EM for k-clustering

Data: n the number of features, *max_iterations, wp.*
Result: A set of selected features xbest, clusters with maximum of likelihood and best accuracy rate.

```
1 : Start with all features;
2 : Apply EM on the current data;
3 : Evaluate the quality of x noted f(x)
4 : For (i = 1 to max_iterations) do
5 : r ←random number between 0 and1.
6 : if ( r < wp) then /*Step 1*/
7 : Generate the neighborhood solutions of x;
8 : Apply EM on the current data; Evaluate the quality of each neighbor x f(x)
9 : xnewi ← pick a best neighbor solution
10 : else /*Step 2 */
11 : xnewi ← a random neighbor solution
12 : Apply EM on the current data; Evaluate the quality of each neighbor x f(x)
13: endIf
14: if ( f(xnewi ) < f(x)) then : x=xnewi; save best accuracy; endIf
15: endfor
```

5. Experiments

In this section, we start with the description of the considered financial datasets used in this study. Then, we give some numerical results found by the proposed approach. The code sources are written in Java. All experiments were run on an Intel Core(TM) i5-2217U CPU@1.70 GHz with 6 GB of RAM under Windows 8 64 bits, processor x64.

5.1 The considred datasets

We evaluate our method on five financial datasets which are: Australian, German, Japanese, Polish and Indian Qualitative Bankruptcy datasets available on UCI (University of California at Irvine) Machine Learning Repository[1]. Table 1 gives details about the considered datasets.

Table 1: Description of the datasets used in the study

Dataset	#Loans	#Good Loans	#Bad Loans	# features
Australian	690	307	383	15
Japanese	690	307	383	16
German	1000	700	300	21
Polish	5910	410	5500	65
Indian	250	143	107	7

5.2 The Numerical results

Tables 2 to 6 give the numerical results found by both the EM method and the method combining SLS with EM. For each method, we give the clustered instances, the number of features and the Log-likelihood. Also, we give TN, TP, FP, FN values and the accuracy rate. The best results are in bold font.

From the experiments, we can see that the proposed method succeeds in finding good results. The SLS succeeds in reducing the number of the features and in improving highly the accuracy rate for all the considered datasets. The proposed method succeeds in splitting efficiently the data into two main groups which are similar to the predefined classes. It permits to enhance the goodness of the clustering. This is done for all the considered datasets. For example, for the Australian dataset (see Table 2), when we used EM alone with all the features, we obtained a TN= 86, TP=98, FP=297 and FN=209. The Log-likelihood was -10.30708. The accuracy rate is 26.68%. However with the proposed method SLS with EM, the number of the considered features is reduced to 11. The TN = 356, TP=244, FP=63 and FN is reduced to 27. The Log-likelihood is improved to **-8.93208**. Also the accuracy rate is improved to **86.96%.** The same remark is noted with the German dataset (see Table 3). With EM, we obtained a TN= 158, TP=241, FP=142 and FN=459. The Log-likelihood was -13.71043. The results are improved when we added the SLS method to EM, we obtained a TN= 68, TP=614, FP=86 and FN=232. The Log-likelihood is **-8.91616**. The accuracy rate was 39.90% with EM while when we executed the SLS with EM, the accuracy rate is improved to **68.2%** and with a reduced number of features equals to **13**. A nice improvement is noted with the Japanese dataset (see Table 4). The accuracy rate was 26.67% and the Log-likelihood was -10.88483 with EM while SLS with EM improved the Log-likelihood to **-9.47562,** the accuracy rate is improved to **86.96%** and with a reduced number of features equals to **12.** Also the TN, TP, FP and FN values are improved. For the Polish dataset (see Table 5), the accuracy rate was 79.86% when using the EM clustering. The accuracy rate is enhanced when we use

1 UCI web site: https://archive.ics.uci.edu/ml/datasets

the SLS with EM. The accuracy rate is increased to **83.59%** with a reduced number of features equals to **56**. The Log-likelihood is enhanced from -38.42108 (with EM) to **-33.14255** (with SLS+EM). Also all the TN, TP, FP and FN values are enhanced when combined SL S with EM clustering. For the Indian dataset (see Table 6), the accuracy rate was 99.2% when using only EM. We remark a slight improvement in favor of our method. The accuracy rate is improved to **99.6%** when using SLS combined with EM with only **5** selected features. The TN and the FN values are enhanced when using SLS. Also the Log-likelihood is enhanced from -5.56554 when using only EM to **-4.58554** when we added SLS to EM.

Table 2: EM versus EM with SLS on Australian dataset (**690 instances)**

Method	**TN**	**TP**	**FP**	**FN**	**Accuracy rate**
EM	86	98	297	209	26.68%
EM with SLS	**356**	**244**	**63**	**27**	**86.96%**
	Clustered Instances			**Log likelihood**	
EM	0: 395 (57%) / 1 : 295 (43%)			-10.30708	
EM with SLS	0 : 271 (39%) / 1 : 419 (61%)			**-8.93208**	
EM	All features (15)				
EM with SLS	**11 selected features**	The best vector solution: 0 0 1 1 1 1 1 1 0 1 1 1 1 1			

Table 3: EM versus EM with SLS on German dataset (**1000 instances)**

Method	**TN**	**TP**	**FP**	**FN**	**Accuracy rate**
EM	158	241	142	459	39.90%
EM with SLS	**68**	**614**	**86**	**232**	**68.2%**
	Clustered Instances			**Log-likelihood**	
EM	0 383 (38%)/ 1 617 (62%)			-13.71043	
EM with SLS	0 846 (85%)/ 1 154 (15%)			**-8.91616**	
EM	All features (21)				
EM with SLS	**13 selected features**	The best vector solution 1 0 1 0 0 1 0 1 1 0 1 1 0 1 1 0 1 1 1 1			

Table 4: EM versus EM with SLS on Japanese dataset (**690 instances)**

Method	**TN**	**TP**	**FP**	**FN**	**Accuracy rate**
EM	86	:98	297	209	26.67%
EM with SLS	**353**	**247**	**60**	**30**	**86.96%**
	Clustered Instances			**Log-likelihood**	
EM	0 395 (57%) / 1 295 (43%)			-10.88483	
EM with SLS	0 277 (40%)/ 1 413 (60%)			**-9.47562**	
EM	All features (15)				
EM with SLS	**12 Selected features**	The best vector solution: 1 1 0 1 0 1 1 1 1 0 1 1 1 1 1			

Table 5: EM versus EM with SLS on Polish dataset (**5910 instances)**

Method	**TN**	**TP**	**FP**	**FN**	**Accuracy rate**
EM	292	4428	118	1072	79.86%
EM with SLS	**259**	**4681**	**819**	**151**	**83.59%**
	Clustered Instances			**Log-likelihood**	
EM	0 4546 (77%)/ 1 1364 (23%)			-38.42108	
EM with SLS	0 4832 (82%)/ 1 1078 (18%)			**-33.14255**	
Selected features					
EM	All features (65)				
EM with SLS	**56 selected features**	The best vector solution: 1 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1			

Table 6: EM versus EM with SLS on Indian. (**250 instances**)

Method	**TN**	**TP**	**FP**	**FN**	**Accuracy rate**
EM	142	106	1	1	99.2%
EM with SLS	**143**	**106**	**1**	**0**	**99.6%**
	Clustered Instances			**Log-likelihood**	
EM	0 107 (43%)/ 1 143 (57%)			-5.56554	
EM with SLS	0 106 (42%) / 1 144 (58%)			**-4.58554**	
EM	All features (7)				
EM with SLS	**5 selected features**	The best vector solution: 1 0 1 1 1 1			

In overall and according to the numerical results, we can say that the proposed k-clustering technique succeeds in improving the goodness of the obtained clusters. For German, Australian, Japanese, Indian and Polish datasets, SLS with EM gives high quality results compared to the pure EM. This is due to the SLS search method that permits to select a set of potential features to use in the clustering task which improve the accuracy rate.

6. Conclusion

This paper proposed a stochastic local search based feature selection method to improve the performance of the EM clustering. The method is applied to credit risk assessment and validated on several financial datasets. The role of SLS is to select a set of potential features to be used in the clustering task. The role of EM is to split data into two main clusters: bad and good. The numerical results are promising and show the importance of the feature selection in data clustering for credit risk. When SLS is used with EM and validated on the considered datasets, the results are improved. The TN, TP, FP and FN values are improved. Also the *Log-likelihood* and the accuracy rate are enhanced. The SLS improved highly the data clustering for all the considered datasets. It would be nice to validate the method on other datasets. This will be the aim of a future work.

References

[1] H. Abdou and J. Pointon (2011), "Credit scoring, statistical techniques and evaluation criteria: A review of the literature", *in Intelligent Systems in Accounting, Finance and Management*, Vol. 18, No. 2-3, pp. 59-88.

[2] J. Abelln, C.J. Mantas, (2014). "Improving experimental studies about ensembles of classifiers for bankruptcy prediction and credit scoring", *in Expert Systems with Applications*, 41, pp. 3825-3830.

[3] T. Bellotti, J. Crook, (2009). "Support vector machines for credit scoring and discovery of significant features". *In Expert Systems with Applications*, 2009, 36, pp.3302-3308.

[4] D. Boughaci (2013: “Metaheuristic Approaches for the Winner Determination Problem in Combinatorial Auction”. *In Artificial Intelligence, Evolutionary Computing and Metaheuristics* 2013, book series (SCI, volume 427), pp. 775-791.

[5] D. Boughaci and A.A.K. Alkhawaldeh (2018), "Three local search based methods for feature selection in credit scoring", *in Vietnam J. Computer Science* Vol 5, N°2, pp. 107-121 (2018).

[6] D. Boughaci and A.A.K. Alkhawaldeh (2018), "A new variable selection method applied to credit scoring", *in Algorithmic Finance,* Vol 7(1-2), pp. 43-52 (2018).

[7] L.Breiman, J.Friedman, R.Olshen, and C.Stone,(1984). "Classification and Regression Trees". Belmont, CA: Wadsworth, 1984.

[8] A.P.Dempster, N. M.Laird, and D.B.Rubin, (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society.* Series B (Methodological) 39 (1): 1–38. JSTOR 2984875.MR0501537.

[9] V. Desay, J.N. Crook, G.A. Overstreet. (1996). "A Comparison of Neural Networks and Linear Scoring Models in the Credit Union Environment", *in European Journal of Operational Research*, 95, 1996, pp. 24-37

[10] N. Friedman, D. Geiger, and M. Goldszmidt, (1997). "Bayesian Network Classifiers", *in Machine Learning,* vol. 29, pp. 131-163.

[11] F. Gonzales, F., F. Haas, R. Johannes, M. Persson, L. Toledo, R. Violi, C. Zins, M. Wieland, (2004), "Market dynamics associated with credit ratings: a literature review", *Banque de France in Financial Stability Review*, 4: 53 -76.

[12] D.J. Hand, W.E. Henley, (1997). "Statistical Classification Methods in Consumer Credit Scoring", *in Journal of the Royal Statistical Society*, Series A (Statistics in Society), vol. 160, pp. 523-541.

[13] Ho, Tin Kam (1995). "Random Decision Forests". *In Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, 1416 August 1995. pp. 278-282.