Presentation of graduation internship

Implementing Machine Learning Methods In Credit Risk Management

Inter: Zinan ZHOU
Supervisor: Dritan NACE

Introduction and state of art

- Build a model that borrowers can use to help make the best financial decisions.
- Classify clients as good clients and bad clients
- Machine Learning models

Techniques used for credit risk analysis could be categorized:

- Statistical methods:
  Logistic regression, classification tree, etc
- Artificial Intelligence techniques:
  Support Vector Machine, artificial neural networks, etc
- Hybrid approaches and Ensemble methods

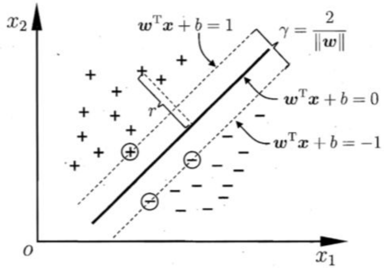
Methods	German Dataset			Australian Dataset		
	OA (%)	Se(%)	Sp(%)	OA(%)	Se(%)	Sp(%)
LDA	72.00	72.43	71.00	85.80	92.50	80.42
QDA	68.00	67.71	68.67	80.59	66.48	91.91
LogR	76.40	88.14	49.00	86.53	88.28	85.12
DT	71.80	79.57	53.67	82.18	80.41	83.56
k-NN	69.90	89.85	23.33	69.13	54.39	80.94
DSlssvm	77.10	88.86	49.67	86.96	89.25	85.12

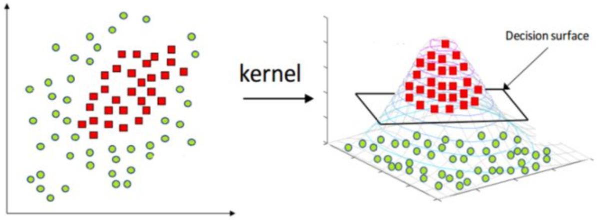
Table 1. Performance comparisons of different classifiers*

* Bio-inspired credit risk analysis: Computational intelligence with support vector machines. Lean Yu, Kin Keung Lai

Principle of SVM

Support Vector Machine





Kernel Trick converts non-linear classification to linear classification by shifting the lower dimension space to higher dimensional space.

SVM decision function:

$$f(x)=\operatorname{sign}(w\cdot\Phi(x)+b)=\operatorname{sign}\left(\sum_{i=1}^{l}\alpha_i y_i K\left(x_i,x\right)+b\right)$$

a hyperplan → SVM → SVM (soft margin)

$$w\cdot\Phi(x)+b=0$$

$$\underset{\mathbf{w},b}{\text{minimize}}\ \frac{1}{2}\ \|\ \mathbf{w}\|^2$$
$$\text{subject to } y_i(\mathbf{w}\cdot\mathbf{x}_i)+b\geq 1,$$
$$i=1,\ldots,m$$

$$\underset{\mathbf{w},b,\zeta}{\text{minimize}}\ \frac{1}{2}\ \|\ \mathbf{w}\ \|^2+C\sum_{i=1}^{m}\zeta_i$$
$$\text{subject to } y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1-\zeta_i$$
$$\zeta_i\geq 0\quad \text{for any } i=1,\ldots,m$$

Model Selection

- Data processing
  - Principal Components Analysis
- Model selection
  - SVM and SVM based models
- Model evaluation
  - Bad precision

  correctly predicted as bad / classified as bad
  - Good precision
  - Total accuracy
  - AUC

Data imbalance problem :
- Over sampling and under sampling methods

Our results with advanced SVM methods :

	Bad precision	Good precision	Total accuracy	AUC
SVM	0.294	0.887	0.855	0.59
Fuzzy SVM	**0.625**	0.716	0.711	**0.67**
Bilateral Fuzzy SVM	0.549	0.781	0.769	0.665
Least Square Fuzzy SVM	0.37	**0.891**	**0.866**	0.63
Weighted LSSVM	0.412	0.852	0.828	0.632
LS Bilateral FuzzySVM	0.444	0.845	0.826	0.645

- ❖ CVXOPT: Python Software for Convex Optimization
- ❖ sklearn.svm.SVC

- ❖ The data are retrieved from the loan inventory of a bank out of France.

Model Selection

- Good clients : Bad clients = 20 : 1
- Data extremely imbalancing
- Random Forest models
  - Balanced Random Forest (BRF)
  - Improved Balanced Random Forest

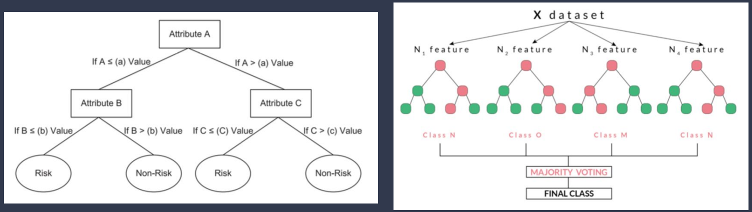


Fig 2. An exemple of decision tree and an exemple of random forest

* An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Iain Brown and Christophe Mues. 2012

	Bad precision	Good precision	Total accuracy	AUC
Fuzzy SVM	0.625	0.716	0.711	0.67
Balanced Random Forest (BRF)	0.838	0.764	0.768	0.801
Improved BRF	**0.908**	0.701	0.708	0.799

❖ imblearn: imbalanced learn

Algorithm of Improved Balanced Random Forest model:

1. Set the threshold :
   0 < threshold_low,threshold_high < 1
2. Training the model and consider the predicted probability [threshold_low,threshold_high] as uncertain.
3. Put the uncertain part into the next BRF model for training, get new predicted probability.
4. Repeat step 2,3.
5. Output: the average of the predicted probability Prob_final.
   Prob_final > 0.5: Good clients; Prob_final < 0.5: Bad clients

Conclution

1.Programming 5 algorithms based on SVM in Python :

- Fuzzy SVM, Bilateral Fuzzy SVM,
- Least Square Fuzzy SVM,
- Weighted Least Square SVM,
- Least Square Bilateral Fuzzy SVM.

2. Data imbalancing degrades SVM model performance. Verifying that Random Forest models perform better when the data is extremely imbalanced.

3. We improved the Balanced Random Forest model, which greatly improved the accuracy of minority class.

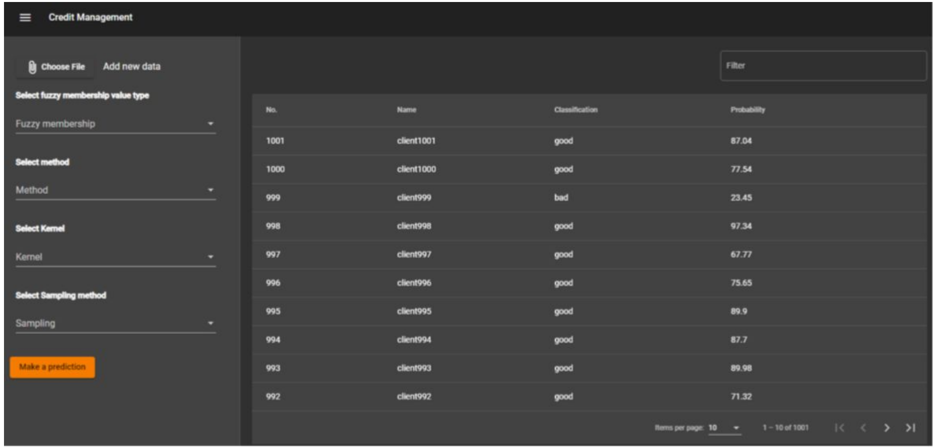


Fig. Frontend of the application

4. Save the model and cooperate with colleagues to complete the user interface.

5. Solutions to classification problems can also be applied to medical and other fields.

Thank you for your attention !