International Journal of Computer Sciences and Engineering
Open Access
Research/ Review/Survey Paper Volume-5, Issue-8 E-ISSN: 2347-2693

Important Fund Client Classification Based On Proportion SVM

Reshma R Krishnan[1] , Thomas George K(Assistant Professor)[2]

[1]Computer science and Engineering, Jyothi Engineering College, Thrissur, India
[2]Computer science and Engineering, Jyothi Engineering College, Thrissur, India
e-mail: rreshmakrishnan@gmail.com, thomas@jecc.ac.in

Abstract— Clients identification for a firm has an important role in making profit. Important client's identification helps the authority to provide sufficient services which helps to maintain the customers in the enterprise. In banking area identity of the client must be a confidential one. So we cannot identify that if the person is important client or not. Sometimes we only get the proportion information about the identity. In this paper we are using this proportional information for identifying the important customers. We are using proportion svm for inferring the individual label from the label proportions. This approach helps to eliminate the constraints on data. The proportion svm model uses two algorithms for improving its efficiency. Simple alternating optimization and algorithm based on convex relaxation. It is more efficient than other label proportion methods like inverse calibration.

Keywords: proportion SVM, Inverse Calibration, label proportions

I. INTRODUCTION

Clients are important factors of a firm. Understanding various clients helps to improve the performance and efficiency. It also helps to improve the profit. We can classify the clients into various levels based on some attributes and from this classification important clients are identified. Based on the significance of each customer company can provide its services and confined resources.

But in banking, security has a vital role so that identity of the client is not visible. However proportions of important clients can be inferred from the dataset based on knowledge. This paper intended to get individuals identity from this proportion information. For this, it proposes proportion svm method which uses two algorithms for efficiency. The training set provided as in the form of bags, which contains only the proportions of the samples. The goal is to find or estimate the individual labels. It treats anonymous label instances as inherent variables which avoids making of constraints on data. It addresses the privacy issue. The efficiency of the method is increased with respect to increase in the bag size. The bags are independent to each other. The learning framework consists of groups of unlabelled instances and only the proportions are given. This can be applicable in other areas like e-commerce and spam filtering.

II. RELATED WORK

Nowadays many researches are conducting for identifying important customers. Gianfranco Chicco [1] studied about different clustering techniques for the classification of electricity customers. It mainly focused on unsupervised clustering algorithms. It also uses self organizing maps to group customers into clusters. The paper proposes a different data reduction technique in order to reduce the size of the input. The paper concludes that the clustering algorithm modified follow-the-leader and hierarchical clustering are comparatively effective algorithms.

Wu R S, Chou P H [2] proposes a soft clustering approach for customer segmentation of multiple category data in e-commerce. The online users are segmented into different categories. The proposed method uses a latent mixed-class membership approach. It also provides high scalability. The segmentation helps to get knowledge about the users which improves the customer relationships. So that new marketing schemes can develop in order to achieve user's expectations. For constructing segments a new methodology that inferred from the latent Dirichlet allocation model has been used. The paper works on multi-category dataset. To calculating membership value for each customer, probabilistic assumptions were formulated. The patterns are represented by matrices. The similarity of clients is spectaculated from the interactions. Main disadvantage of the soft clustering approach is large number of iterations are needed for the computation. It is also difficult to set the number of classes and selecting cut off values for the membership classes. The percentage of cut off values is affected by similarity function.

Lopez J J, Aguado J A, Martin F suggests a classification method for electric customers which uses a Hopfield recurrent ANN [3].The paper proposing accurate and efficient classification of customers which helps to develop

new methodology for tariff calculation and pricing. The paper divides the users into different clusters based on recurrent artificial neural network. For data reduction different filtering techniques are used. The efficiency of the method is evaluated using characterization indexes. The paper compares different data reduction techniques and examines its results. These data reduction techniques include hour load profile, form factors method, harmonic analysis, time frequency analysis and the principal component analysis. All these methods make Hopfield ANN sustained and having a minimum value. The PCA provides better result when compared to another method based on the computation time. It only considers three parameters in the load curve.

Chen Y L, Hsu C L, Chou S C [4] proposes construction of multi-valued and multi-labeled decision tree. In recent days, different application needs multi valued and multi labeled data. So the requirement for a novel classification algorithm which deals with multi valued data is increases. This algorithm defines how to select a node which is used for splitting the tree and when to stop splitting. The algorithm aims to build a classifier tree.

Hu T L, Sheu J B [5] proposes a fuzzy-based customer classification. It mainly focused on business logistical markets. The method helps to group customers before performing fleet routing. The clustering technique mainly consists of three steps: first step includes binary transformation. It includes converting decision variables into binary data. The second step involves creating a fuzzy correlation matrix. This matrix represents the correlation between pair of customers. Third step includes grouping the customers. Each group contains customers with high similarities in terms of their attributes. It makes major improvements in demand and supply sides. It reduces the complexity and increases the efficiency so that pre trip customer classification can do effectively. Fuzzy clustering is an efficient technique which improves both data compression and data categorization. It also classifies clients in terms of their demand variables.

III. METHODOLOGY

The main goal of the proposed system is to identify significant clients in a commercial bank by using proportion SVM method. The Proportion SVM method uses learning with label proportions. That is, the training dataset is divided into different bags. Each bag contains the proportion of the labels. This is because of the privacy issues of the customers. Since in banking area, information about the individual kept confidentially we cannot identify individual labels. The proportion SVM is an efficient method to predict the individual label from the label proportions.

A.PROPORTION SVM

Proportion SVM uses a binary learning setting. It takes training dataset as k bags,

$\{x_i|i£ B_k\}^k_{k=1}$, $U_{k=1}{}^k B_k = \{1,\ldots N\}$ (1)

Assuming that bags are independent and k^{th} bag contains label proportions p_k:

For all k= {1,..N},

$|\{i|i£B_k, y_i^* = 1\}| / |B_k|$ (2)

In which $y_i^* = \{1,-1\}$ denotes the unknown ground truth label of xi, We use $f(x) = sign(w^T \psi(x)+b)$ for predicting the binary label of an instance x, where ψ (.) is a map of the input data.

Here $y=(y_1,\ldots,y_N)^T$ in which y_i£ {1,-1}.

Therefore proportion of the kth bag is given by

$\tilde{p_k}(y) = |\{i|i£B_k, y_i^* = 1\}| / |B_k|$

$= \sum_i £_{Bk}\, y_i / 2|B_k| + 1/2$ (3)

Under the large margin framework

$Min_{y,\, w,b}\ 1/2w^Tw + C\sum_{i=1}{}^N L(yi; w^T \psi(xi)+b)$

$+Cp\sum Lp(\tilde{p}_k(y), pk)$ (4)

In which L(.) ≥0 is the loss function. Lp(.) ≥ 0 is a function to penalize the difference between the true label proportion and the estimated label proportion based on y. The aim is to optimize the labels y and the model parameters w and b.It uses different loss functions.[7]

The proportion SVM uses a simple alternating optimization method and a convex relaxation method for improving the efficiency.

B. ALTERNATING PROPORTION SVM METHOD

For a fixed y, the optimization of (4) with respect to w and b became a classic SVM problem. This algorithm first fix the value of y to solve w and b iteratively until the decrease of the objective is smaller than a threshold. This alternating values fixing leads to convergence since the objective function is non-increasing. Main problem is the chances for local solutions.

C. CONVEX PROPORTION SVM METHOD

This method converts the objective function to a convex function of M: $=yy^T$.Then solution space is relaxed to its convex hull; causes to a convex optimization problem of M.The algorithm solve the relaxed problem.

IV. CONCLUSION

Client's identification has an important role in the success of an organization. Identifying whether a customer is important or not helps the authority to provide limited service

distribution. It helps to maintain the important customers in the organization. Sensitive information about the clients must be kept confidentially especially in the banks and hospitals. Due to these security problems, specific label of commercial bank clients are not available.Inorder to address the problem, this paper proposes proportion svm method which predict the label of individual client. The data is divided into different bags which are independent. Each bag contains only the proportion information.

REFERENCES

[1] Chicco G, Napoli R, Piglione F. "*Comparisons among clustering techniques for electricity customer classification"*,Power Systems, IEEE Transactions on **2006**, **21(2)**: **933-940**.

[2] Wu R S, Chou P H, "*Customer segmentation of multiple category data in e-commerce using a soft-clustering Approach"*, Electronic Commerce Research and Applications, **2011**, **10(3)**: **331-341**.

[3] L´opez J J, Aguado J A, Martin F, "*Electric customer classification using Nopfield recurrent ANN*",EEM **2008.** 5th International Conference on European.IEEE, **2008: 1-6**

[4] Chen Y L, Hsu C L, Chou S C.,"*Constructing a multi-valued and multi-labeled decision tree"*, Expert Systems with Applications, **2003, 25(2): 199-209**

[5] Hu T L, Sheu J B., "*A fuzzy-based customer classification method for demand-responsive logistical distribution Operations"*, Fuzzy Sets and Systems, **2003**, **139(2): 431-450**.

[6] Shin H W, Sohn S Y, "*Segmentation of stock trading clients according to potential value"*, Expert systems with applications, **2004**, **27(1): 27-33**.

[7] Felix X.Yu,DongLiu,SanjivKumar,Tony Jebara, Shih-Fu Chang,"*αSVM for Learning with Label Proportions"*, Columbia University, New York, NY **10027**,Google Research.

[8] Lai K T, Yu F X, Chen M S,"*Video event detection by inferring temporal instance labels"*,Computer Vision and Pattern Recognition (CVPR), **2014** IEEE Conference on. IEEE, **2014**: **2251-2258**.

[9] M. Stolpe, K. Morik,"*Learning from label proportions by optimizing cluster model selection"*,Machine Learning and Knowledge Discovery in Databases,Springer, **2011**, pp**. 349{364**.

[10] S. Rueping, "*Svm classier estimation from group probabilities*", Proceedings of the 27th International Conference on Machine Learning(ICML-**10**), **2010**, pp. **911{918**.

