IJCAES
ISSN: 2231-4946
Volume V, Special Issue, January 2015

International Journal of Computer Applications in Engineering Sciences
Special Issue on International Conference on Computing, Communication and Sensor Network (CCSN-2014)
www.caesjournals.org

An Intelligent prediction model for identifying target customers

Prathima Guruprasad[1], Guruprasad K S Rao[2]

[1]Department of Computer Science and Engineering, Assistant Professor, VKIT, Bangalore, Karnataka, India[2]Principal Consultant, Mindtree Limited, Bangalore, Karnataka, India
[1,2]University of Mysore (Research Centre: Nitte Research and Education Academy)

[1]prathimaguru18@gmail.com
[2]guruprasad_rao@mindtree.com

***Abstract*—Soft computing methodologies are characterized by the use of inexact solutions to computationally challenging tasks. One of the typical applications of Soft computing techniques is in finding the relation of input attributes to a target item. In this paper we intend to find the significant attributes from larger attribute list to find its impact on a target item. Using Support vector machine (SVM), the system has achieved a good prediction accuracy of 98.06%. This would aid the marketing team to predict their target customersby focusing on relevant components in data to plan an expert marketing strategy.**

***Keywords*—Soft Computing, Support Vector Machines, Chi-square, Feature Analysis, Marketing Strategy**

I. INTRODUCTION

This paper provides a critical analysis of a large multidimensional data set with the intent of identifying valid, new, previously unknown, decisive patterns in data. Soft computing is used to identify relationships among a set of items in a database. These relationships are not based on inherent properties of the data themselves (as with functional dependencies), but rather based on co-occurrence of the data items. In this process we discover a set of association rules at multiple levels of abstraction from the relevant set(s) of data in a database. The main constitutes of soft computing include fuzzy logic, neural networks, Support Vector Machines, genetic algorithms and rough sets. Each of them contributes a distinct methodology for addressing problems in its domain. The actual soft computing task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown but interesting patterns such as groups of data records. It aids to check unusual records or detects anomaly in data and serves to spell out dependencies which is otherwise called as association rule mining. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the soft computing step would identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system.

Factor analysis is a key constituent of Multivariate analysis which is used to uncover the latent structure of a set of variables. It reduces the attribute space from a larger number of variables to a smaller number of factors. Factor analysis originated a century ago with Charles Spearman's attempts to show that a wide variety of mental tests could be explained by a single underlying intelligence factor. Some of its applications include:

- To reduce a large number of variables to a smaller number for data modeling
- To select a subset of variables from a larger set based on which original variables have the highest correlations with some other factors.

II. PROBLEM DESCRIPTION

Vendors are constantly on the move to identify their target customers. That would aid a proper mapping between the supplier and the consumer and satisfy ones need or want. Such segments provide constant business which forms the biggest revenue for organizations. Not identifying the right customer segments affects organizations very badly and results in dropping revenues. Organizations also would lose out to competitors who would not only identify but exceed the expectations of customers currently in some other organization's segment. Hence it would expose risk losing business to competitors.

Given a large database and a multitude set of attributes, the goal of this study is to classify the dataset using a subset of features which have the largest "impact". These would have their own order of importance on how they affect the target item. Using a supervised technique we partition the survey data to a training and test set. The training set is used to obtain the key pattern in the data with the subset of attributes. We apply this to the test data to find out the accuracy of recommendation by applying the soft computing paradigm. This is compared with another method of factor analysis to reflect the accuracy of prediction.

III. ARCHITECTURE OF PREDICTOR MODEL

The predictor uses a THREE step model involving the following as shown in fig 1:

a. Attribute Evaluation: We use Chi-square attribute evaluation scheme.
b. Ranking and selection of attributes: Order used is based on the degree of significance of each attribute. The top 21 attributes are selected.
c. Classify with the use of Support Vector Machine or Linear Regression.

Chi-square [1] is a statistical test commonly used to compare actual observed data with expected data derived according to a specific hypothesis. This test provides the "goodness to fit" between the observed and expected data. The test results in either of the following two cases:

a. We **reject** the null hypothesis and consider that there **IS** association between the test input attribute and the target item since the Critical value is < test value
b. We **accept** the null hypothesis and consider that there **isNO** association between the test variable and target variable since Critical value is > test value. In other words it means that there is NO significant impact of the input attribute being tested to the target item.

$$ChiSquare-Statistic:$$
$$\chi = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)}{E_i}\right]^2 \qquad (1)$$

Where: O_i = The observed frequency , E_i = The expected frequency and *K = Number of Instances*

The critical values are found from the standard Chi-square distribution tables with a degree of freedom that is computed against : (r–1) where r = # of rows obtained after the observed instances are grouped to specific categories before analysis.

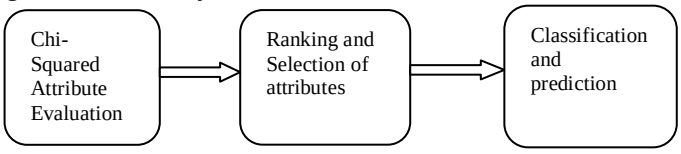

Figure 1: Architecture of Predictor Model

A. Data Set

The dataset provided in the CoIL Challenge 2000 data mining competition [2, 3] is used here. Our training data set has 5822 customer records with each record having multi-dimensional attribute of 85 containing socio-demographic data. The target item is 86th column which is a number of mobile home policies of a specific type of Insurance policy.

B. Feature Analysis

The number of dimensions and the size of the data set prompt us to look for ways to analyse and remove insignificant dimensions. Using Chi-Square analysis, the key attribute having the largest impact is indicated in Fig 2.

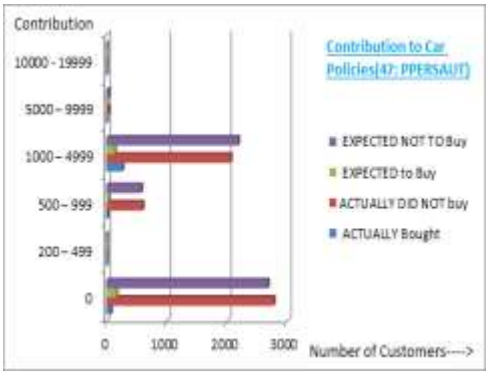

Fig. 2 Actual and Expected value of Contribution value of an input attribute to the number of customers

C. Feature Selection

The Chi square value for this feature is indicated in Fig 3 below. The Critical value for 5 % level of significance against the stated degree of freedom is also shown in the same figure. We notice that the 4th category indicates the largest impact with the overall impact being the highest on target item.

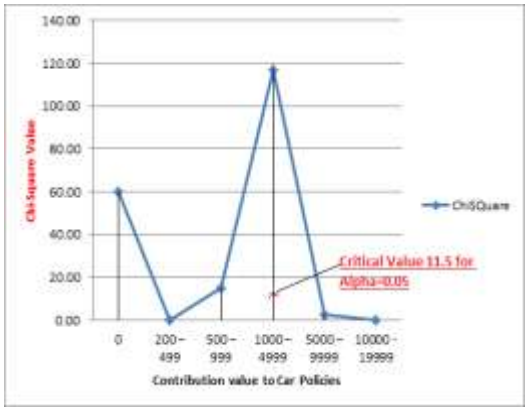

Fig. 3: Chi-Square values for various categories of contribution value to the Target item.

The Chi-square is similarly computed for each feature in the order or rank of impact. The results of the feature are tabulated in table 1 below.

TABLE 1 : RANK AND LABEL OF INPUT ATTRIBUTES

Rank	Attribute Label	Attribute Number
1	PPERSAUT	47
2	APERSAUT	68
3	MOSHOOFD	5
4	MKOOPKLA	43
5	APLEZIER	82
6	PWAPART	44
7	MINKGEM	42
8	AWAPART	65
9	PBRAND	59
10	PPLEZIER	61
11	MOSTYPE	1
12	MOPLLAAG	18

13	MINKM30	37
14	MSKA	25
15	MHHUUR	30
16	MHKOOP	31
17	MOPLHOOG	16
18	ABYSTAND	85
19	MAUT0	34
20	PBYSTAND	64
21	ABRAND	80

D. Classification

Support Vector Machine is a learning algorithm used in soft computing for classification problems such as data mining, text categorization, handwritten character recognition, image classification and Facial expression classification. When we know nothing about how we classified it, or we don't know the rules used for classification, when a new data comes, SVM can PREDICT which set the data should belong to.

Libsvm is a library for support vector machine [4]. Libsvm is used with Radial Basis Kernel function (RBF) and is denoted as follows:

$$K(x_i, x_j) = \exp(-\gamma \| x_i - x_j \|^2) \quad (2)$$

The model obtained after training 75% of the customer instances to the SVM is indicated in Table 2. The support vector representation is show in Fig 4:

TABLE 2 : SVM PARAMETERS

SVM Type	Kernel Type	Gamma	rho	Label/ Number of SV	
c_svc	rbf	0.0117647	-0.996327	0	1
				601	83

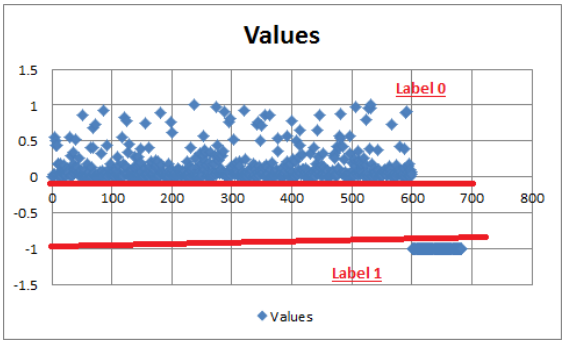

Fig. 4: Support vector representation for two classes with raw data.

IV. PERFORMANCE ANALYSIS

For Raw data, the classification performance with SVM is 81.07%. However, for ranked and selected attributes of 21, the performance increases to 98.06 percent as shown in Table 3 and Fig 5. We compare this with linear regression which provides a classification accuracy of 80%. The results of linear regression are shown in Fig 6. This prediction model can applied to huge data sets in fields of medical sciences [5, 6] and finance.

TABLE 3 : SVM CLASSIFIER

Category of Test	# of Iterations	rho	Number of Support Vectors	Classification Accuracy	Classification
1	4161	-0.86661	1696	95.91%	4241/4422
2	4065	-0.87925	1609	98.36%	1378/1401
3	4274	-0.86698	1705	99.93%	1400/1401

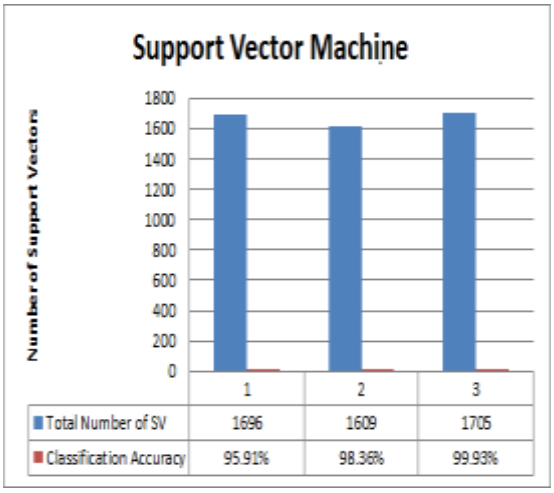

Fig. 5:Performance of Support vector machines for the Top 21 attributes

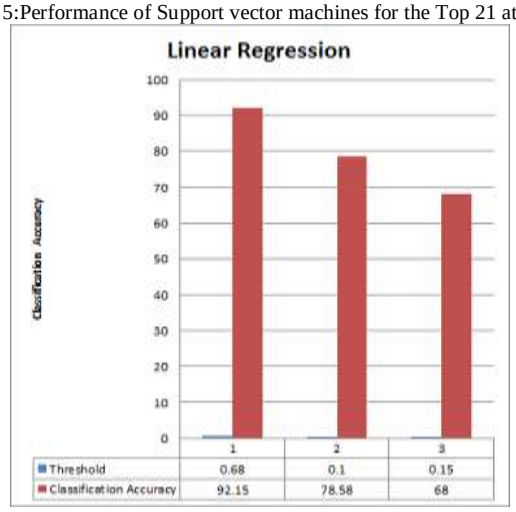

Fig. 6:Performance results of linear regression model for top rated attributes

V. CONCLUSION

This Paper provides an intelligent approach to resolving a marketing problem with a model that combines the aspects of feature analysis, ranking, selection and classification techniques to improve the prediction accuracy. Feature analysis and selection is achieved using Chi-square statistics.The resulting model is ranked in the order of its impact to the target item. This work demonstrates the effectiveness of using the combination of Chi-square along with one of the classifiers such as Support Vector Machine or Linear Regression. The SVM shows improved performance in accuracy, especially with reduced feature

selection.The prediction model is validated with the test instances with an average classifier accuracy of 98.06% as against the prediction of raw data being 81.07%. A methodology can be developed to select the relevant number of attributes automatically based on user defined threshold using any feature selection algorithm. This prediction model can be applied to data sets in other engineering applications such as medicine, finance and enterprise web data.

REFERENCES

[1] Les Oakshott, " Essential Quantitative Methods: For Business, Management and Finance", Fourth Edition 2009, Palgrave Macmillan

[2] Han J Kamber M.: Data mining concepts and techniques.Morgan Kaufmann Publishers(2006)

[3] P. van der Putten and M. van Someren (eds). CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam.http://www.liacs,nl/~putten/library/cc2000/

[4] Chih-Chung Chang and Chih-Jen Lin," LIBSVM: A Library for Support Vector machine", National Taiwan University, Taipei, Taiwan, April 4, 2012

[5] Duvvuri,Venkata , rama,Satya Kumar et al:Management of Filariasis using Prediction Rules derived from Data Mining(2005).

[6] Sellapan,Palaniappan,rafiah Awang:Intelligent heart disease prediction system using data mining techniques.In: international Journal of Computer Science and Network Security vol 8 no 8(2008)