Research Journal of Mathematical and Statistical Sciences ISSN 2320–6047
Vol. **2(2),** 1-6, February **(2014)** *Res. J. Mathematical and Statistical Sci.*

Logistic Regression Classification for Uncertain Data

Abdallah Bashir Musa
College of Mathematics and Computer Science, Hebei University, Baoding 071002, CHINA

Available online at: www.isca.in, www.isca.me
Received 17th October 2013, revised 2nd December 2013, accepted 12th February 2014

Abstract

Logistic regression (LR) is a famous classification technique commonly used in statistics, machine learning, and data mining area of knowledge for learning a response of binary nature. It assumes that the data values are pre-determined precisely, but this is not true for all conditions. Uncertainty data arises in many applications because of data collection methodology as in repeated measures, outdated sources and imprecise measurement as in physical experiments. Studying this uncertainty data becomes area of interest for researchers nowadays. In uncertainty, the value of data item is mostly characterized by a multiple values. So, machine learning techniques are also required to manage an uncertain data. This paper studies the modification of LR technique to handle data with an uncertainty. Statistical inference and theory of probabilities are used to obtain single unbiased estimator that represents the multiple values sufficiently and efficiently. The Maximum Likelihood Estimators (MLE) and the Probabilities Density Function (PDF) are used to capture the uncertainty. Results of the Experiments on UCI data sets demonstrated that the uncertain LR classifier can be constructed successfully, and its accuracy can be improved by taking into consideration the uncertainty information.

Keywords: Logistic Regression (LR), Uncertain data, classification, Maximum Likelihood Estimators (MLE), Probabilities Density Function (PDF).

Introduction

LR model is considered as one of most renowned classification models. LR is appreciated and broadly used in statistics, machine learning, and data classification communities[1,2,3,4]. The advantages of this model include a strong statistical foundation and probabilistic model which helps in analyzing data[5]. It is mostly used in binary classification problems in applied sciences such as medicine, biology and epidemiology. It has been widely applied due to its simplicity and great interpretability[6]. LR has been considered to be an efficient classifier and a powerful prediction method. However, as for most of the machine learning algorithms, it was constructed to handle only the data with certainty where there is a single value in each attribute. Recently uncertainty data are arising in many applications[7,8,9]. Uncertainty can originate due to numerous reasons such as devises precision limitation, data sampling error, and collection data problems as in repeated measures, which it seems to be the common type of data with uncertainty, among others etc, .In uncertainty data information cannot be idyllically represented by a single value. Therefore, it is difficult to achieve satisfactory results when classifying such data without managing uncertainty. The error in the data is generally treated as a random variable with probability distribution[10]. The uncertain attribute value is often represented by an interval with a probability distribution function over this interval[11, 12].

Classification is a classical problem in machine learning and data mining[13]. It is the process of building a model that can describe and predict the class label of data based on features. Many classification algorithms such as support vector machine, decision tree, Bayesian and neural networks used the certainty data for performing the classification tasks. A common way to handle the uncertainty is to represent it by its expected value and treat it as certainty data, so the classification algorithm can be directly applied. However, this method does not effectively utilize important information such as probability function and distribution intervals. Therefore, building classifier based on uncertain data is a great challenge that needs specific classification algorithms.

There are significant studies that proposed classification algorithms to analyze uncertainty data, such as support vector classification with input data uncertainty[14], decision trees for uncertain data[13], a decision tree for classifying uncertain Data[15], naive Bayes classification of uncertain Data[16] a Bayesian classifier for uncertain data[17,] and a neural network for uncertain data classification[10].

To the best of present knowledge, there is no study focus on building LR from uncertainty data, Moreover, the method that propose in this paper is very different from those used in previous studies.

In this paper, the LR classifier algorithm has been extended to handle the uncertainty data The maximum likelihood estimator (MLE)[18,19] which is the best, unbiased, sufficient and efficient estimator is used with the Gaussian and the uniform distributions to represented the multiple values, unlike others studies where average is used as an estimator which is not always good as in

the case of the uniform distribution. The kernel density estimation method[20] which provides an effective representation and good approximation of the density is used. The idea in kernel density estimation is to provide a continuous estimate of the density of data at a given point. The value of density at a given point is estimated as the sum of smoothed values of kernel functions associated with each point in data set. Each kernel function is associated with a kernel width which determines the level of smoothing created by the function. The probabilities density function (PDF) method of the Gaussian distribution is used to capture the uncertainty. The purpose of this study is to construct the LR on the uncertainty data and to investigate the classification accuracy achieved by the average, MLE and PDF.

Logistic regression: Logistic Regression (LR) is a well known statistical classification method for modeling dichotomous (binary) data [1.2.3.4]. Let $x \in R^n$ denote a vector of explanatory or feature variables, and let $y \in \{-1,+1\}$ denote the associated binary class label or outcome. The logistic model is defined as:

$$\text{pr}\,(y / x) = \frac{1}{1+\exp(-y(\beta^{\mathrm{T}} x+\alpha))} = \frac{\exp(y(\beta^{\mathrm{T}} x+\alpha))}{1+\exp(y(\beta^{\mathrm{T}} x+\alpha))} \quad (1)$$

Where Pr(y/x) is the conditional probability of y given $x \in R^n$. The logistic model has parameters $\alpha \in R$ represent the intercept tem and $\beta \in R^n$ represent the weight vector. $\beta^{\mathrm{T}} x+\alpha = 0$ defines a hyperplane in the feature space, on which P(y/x)=0.5. The conditional probability Pr(y|x) is larger than 0.5 if $\beta^{\mathrm{T}} x+\alpha$ has the same sign as y, and less than 0.5 otherwise. Suppose we are given a set of m observed or training data $\{x_i, y_i\}_{i=1}^m$, where $x_i \in R^n$ denote the i-th sample and $y_i \in \{-1,+1\}$ denote the corresponding class label. These m samples are assumed to be independent samples. According to the logistic model, the vector of the conditional probabilities associated of these samples is:

$$\text{pr}\,(\alpha,\beta)_{\mathrm{i}} = \text{p}\,(y_{\mathrm{i}} / x_{\mathrm{i}}) = \frac{\exp y_i(\beta^{\mathrm{T}} x_i+\alpha_i)}{1+\exp y_i(\beta^{\mathrm{T}} x_i+\alpha_i)} \quad i=1,.....,m \quad (2)$$

The likelihood function associated with the samples is $\prod_{i=1}^{m} \text{pr}\,(\alpha,\beta)_{\mathrm{i}}$, and the log likelihood function is:

$$\sum_{i=1}^{m} \log \text{pr}\,(\alpha,\beta)_{\mathrm{i}} = -\sum_{i=1}^{m} f(\beta^{\mathrm{T}} a_i+\alpha y_i) \quad (3)$$

Where $a_i = x_i y_i \in R^n$ and f is the logistic loss function that is:

$$f(z) = \log(1+\exp(-z) \quad (4)$$

Using (4),(3) can be written as

$$\sum_{i=1}^{m} \log \text{pr}\,(\alpha,\beta)_{\mathrm{i}} = -\sum_{i=1}^{m} \log(1+\exp-(\beta^{\mathrm{T}} a_i+\alpha y_i))) \quad (5)$$

The negative of the log likelihood function is called the (empirical) logistic loss, and dividing by m we obtain the average logistic loss,

$$l_{avg}\,(\alpha,\beta)_{\mathrm{i}} = \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp-(\beta^{\mathrm{T}} a_i+\alpha y_i))) \quad (6)$$

The model parameters β and α can be determined by maximum likelihood estimation from the observed examples, by solving the convex optimization problem

$$\text{minimize}\ l_{avg}\,(\alpha,\beta)_{\mathrm{i}} \quad (7)$$

The problem (7) is called the logistic regression problem (LRP). This LRP is a smooth convex optimization problem, and can be solved by a wide variety of methods, such as gradient descent, steepest descent, Newton, quasi-Newton, or conjugate-gradients (CG) methods. In this paper Newton method is used. Once we find maximum likelihood values of α and β, that is, a solution of (7), we can predict the probability of the two possible outcomes. Given a new features vector $x \in R^n$, by using the associated logistic regression model, the logistic regression classifier is formed as:

$$\varphi(x) = \text{sgn}(\beta^T x+\alpha)$$

Where

$$\text{sgn}(z) = \begin{cases} +1 & z>0 \\ -1 & z\le 0 \end{cases} \quad (8)$$

Which picks the more likely outcome, given x, according to the logistic model. When m, the number of training samples is smaller than n, the dimension of the samples, directly solving the logistic regression formulation in (7) is ill-posed and may lead to overfitting. A standard technique to avoid overfitting is regularization.

ℓ1-Regularized Logistic Regression: More recently, ℓ1-regularized logistic regression has received much attention as a promising method for feature selection[21.22]. The ℓ1-regularized logistic regression problem (ℓ1-regularized LRP) can be formulated as:

$$\text{minimize}\ l_{avg}(\alpha,\beta)\ +\lambda\|\beta\|_1 = \frac{1}{m}\sum_{i=1}^{m} \log(1+\exp(-(\beta^T a_i+\alpha y_i))+\lambda\sum_{i=1}^{n}|\beta_i| \quad (9)$$

Where $\|.\|_1$ denote the ℓ1-norm i.e., $\|\beta\|_1 = \sum_{i=1}^{n}|\beta_i|$ and $\lambda > 0$ is a pre-specified regularization parameter. The objective function ℓ1-regularized LRP in (9) is convex, but not differentiable. The main motivation is that ℓ1-regularized LR typically yields a sparse vector β, that is, β typically has

relatively few nonzero coefficients. When $\beta_j = 0$, the associated logistic model does not use the jth component of the feature vector, so sparse β corresponds to a logistic model that uses only a few of the features, that is, components of the feature vector. Indeed, we can think of a sparse β as a selection of the relevant features. The regularization parameter λ roughly controls the number of those nonzero coefficients, with larger λ typically yielding sparser weighted vector β. For solving the ℓ1-regularized LRP (9), generic methods for nondifferentiable convex problems such as the ellipsoid method can be used. In this paper, the method of a preconditioned conjugate gradient (PCG), which know as a best method, is represented in [21] and is used in this paper.

Handling Uncertainty Information: In this section, three techniques for handling uncertain data in LR classification task are proposed; these techniques are average, maximum likelihood estimator (MLE) and the probabilities density function (PDF).

Averaging: The easy intuitive method to deal with uncertainty data in classification is to compute the average of the uncertainty values for each object and used this average as represented value of those values. By using this method all the objects will include a single value. Consequently the data set will be converting to a certainty data and hence the traditional logistic regression classifier' algorithm can be successfully reapplied.

Maximum likelihood estimator (MLE): The maximum likelihood estimation' (MLE) method[18,19] is a powerful and a well-known estimation method in statistics. It estimates the parameters of a statistical model. When the maximum likelihood applied to a data set and given a statistical distribution, it provides estimates for the distribution's parameters. The concept of the maximum likelihood is as follow: for a given a random sample $x_1, x_2, \ldots, x_n$ where $x_i \sim f(x_i, \theta)$, so that the likelihood function is the join density function of the sample i.e

$$L(\theta, x) = f(x_1, \theta) \ldots\ldots\ldots\ldots f(x_n, \theta) \quad (10)$$

By supposing that x_i is identical, independent distribution (iid), the maximum likelihood function can be defined as:

$$L(\theta, x) = \prod_{i=1}^{n} f(x_i, \theta) \quad (11)$$

Most of likelihood functions satisfy the regularity conditions, so maximum likelihood estimator is the solution of the following equation:

$$\frac{\partial}{\partial(\theta)} \log L(\theta, x_1, \ldots . x_n) = \frac{\partial}{\partial(\theta)} \log \prod_{i=1}^{n} f(x_i, \theta) = 0 \quad (12)$$

MLE obtained by using equation (12) is uniformly minimum variance unbiased estimator (UMVUE)[19]. The properties of the maximum-likelihood estimators are: sufficient : include the maximum information that included in the sample, unbiased: the expected value of the MLE is tend to be the estimated parameter θ, the distribution of the MLE tends to the Gaussian distribution with mean θ and covariance matrix equal to the inverse of the Fisher information matrix and efficiency, i.e., it achieves the Cramer–Rao lower bound when the sample size tends to infinity. This means that no asymptotically unbiased estimator has lower asymptotic mean squared error than the MLE, although the MLE is assumed to be consistent, however, it need not be consistent is some condition[23]. More precisely the MLE for the Gaussian and the uniform distribution that we used in this paper are unbiased, consistent, sufficient and efficient. The kernel density estimation is applied her to obtain smoothed values of the estimated function, those values are used to compute the maximum likelihood estimators (MLE) for the distributions.

Distribution based: For the uncertainty model, a feature value is represented by PDF "f_{ij}", not by a single value. A PDF "f_{ij}" could be implemented numerically by storing a set of **s** sample points; $x \in [a_{ij}, b_{ij}]$ with the associated value $f_{ij}(x)$, effectively approximating "f_{ij}" by a discrete distribution with "s" possible values. The PDF is computed for the Gaussian distribution in this interval, i.e. $f_{i,j}(x) = \int_{a_{ij}}^{b_{ij}} \frac{1}{\sqrt{2\sigma\pi}} e \frac{-(x-\mu)^2}{2\sigma^2} dx$, thus the uncertainty data is represented the PDF value in that interval using the s sample points. Using this approach for represent the uncertainty, the amount of information available in the uncertainty data is captured by the PDF, definitely, good classification model can be build by using this rich information.

Methodology

The data sets: The data sets that are used in this study are composed of 5 numerical features. These data sets are downloaded from the UCI Machine Learning repository available at ftptoics.uci.edu/pub/ma chine learning databases; they occur frequently in the literature of the field. Table-1 gives a numerical summary of the data sets.

The experimental set up: The experiments have been performed on the datasets listed in Table-1. These datasets are chosen because they all have numerical attributes. As the original data contain point values without uncertainty, so the following procedures are used to make data values having uncertainty: For the AVG method, the original point value data are used as the expected value and the experiments are performed on the original datasets. For the maximum likelihood estimators (MLE) method, the Gaussian and the uniform distributions are used as the uncertainty models. The original point is used as the mean of the object (μ_j) for the two

distributions, for the Gaussian distribution the standard deviation is set to be $sigma_j = 0.25(\max v - \min v) * w\%$ and for the uniform distribution it set to be $sigma_j = (\max v - \min v)/12 * w\%$, where "maxv "and" minv" are respectively the maximum and minimum values among the whole attributes of the feature respectively, and w is a percentage parameter used to control the uncertainty level of the objects. In this experiment, four values for w are used which are 1%, 5%, 10% and 20%.

A sample of 100 data points is generated for each object for the two distributions using the parameters that have determined previously. The kernel density estimation is applied to those "s" data points to obtain smooth continuous estimate of density of data. The value of density at a given point is estimated as the sum of smoothed values of kernel functions K'h(·) associated with each point in the data set. Each kernel function is associated with a kernel width h called bandwidth which determines the level of smoothing created by the function. For the bandwidth parameter h, a commonly used bandwidth estimation rule called the Silverman approximation rule [20] was used, which suggests setting bandwidth as $width h_j = 1.06 * (sigma_j) * (n_j)^{-(1/5)}$, where "$sigma_j$" is the standard deviation of the object j and "n_j" is the number of the data objects. The maximum likelihood estimators(MLE) is computed from the values that obtained using the kernel function, for the Gaussian distribution the mean is the MLE , for the uniform distribution the maximum ordered value that obtained after reordering the values ascending is represents the MLE. Using these values of MLE, the data sets are again reconverted to certainty data sets, effectively; the traditional LR algorithm is successfully applied. For the probabilities density function (PDF method, the range of the "s" sample data points that is generated previously is noted; $x \in [a_{ij}, b_{ij}]$, the PDF is generated in this interval by considering the Gaussian and the uniform distributions using the "s" sample data points . Consequently, the data set with uncertainty values is transformed into certainty data value represented by the PDF. Similarly, the traditional LR algorithm is applied. The Newton method through l1_logreg package is used to train the LR. For all data sets 10 fold cross-validations method is used.

Table-1
Summary of the data sets

Data set	Data size	Number of variables
Page block	5473	10
Pima diabetes	768	8
Breast cancer(original)	569	30
heart	270	13
ionosphere	351	34

Results and Discussion

The results are obtained using l1_logreg: A large-scale solver for ℓ1-regularized logistic regression problem package version 0.8.2 [24], available at (http://www.stanford.edu/*boyd /l1_logreg/), under matlab (7.8.0347- R2009a) interface, where the Newton method is used for estimation the logistic regression model parameters.

Accuracy improvement: The classification results of applying the maximum likelihood estimators (MLE) for the data sets using Gaussian and the uniform distributions are given in Table -2 and Table -3 respectively. For all data sets a 100 data points (i.e.s=100) are used with the various values for "w". Both the Gaussian and the uniform distributions are applied to all the data sets. Table -4 shows the classification results of applying the average and the probabilities density function (PDF) methods to the data sets considering the Gaussian distribution using the same various levels of "w" and 100 data points for each data set. Because of the limited space of the paper only the results of the accuracy are shown for the classification and the results of the others performance measures are omitted. The best value of the accuracy for each data set is identified by bold font. The relationship between the accuracies of MLE of Gaussian and uniform distributions and the PDF of Gaussian distribution using ("w=20%") is depicted in Figure.1 using the same order in Table-1. It shows that mostly, the accuracies values of the tree approaches of each data set lie closely.

Discussion: From table -2, table -3 and the results of average in the first column in table -4, it is clear that MLE build more accurate classification than the average of the original data for both the Gaussian and the uniform distributions. Using uniform distribution gives better accuracy in 4 out of 5 data sets. From table-4, applying PDF also gives more accurate classification than average considering Gaussian distribution. On the other hand comparing the results of MLE and PDF for Gaussian, both methods almost give the same accuracy for most of data sets. Generally, the results show that the accuracy is improved with increasing the level of "w", especially when applying the method of the PDF, which may mean that more the uncertainty, more the accurate classification. Comparing MLE for Gaussian and uniform distributions with the PDF of the Gaussian distribution, (from figure-1) all the three approaches almost give the same performances, however MLE for the uniform distributions give better accuracy in 4 data sets out of 5, compared to the other two approaches. The experiment has been repeated using varieties of number of sample points ranging from 100 to 1000, keeping the level of "w" constant; no significance improvement in the accuracy was observed. Defiantly, if the data sets that used in this study are originaly uncertainty data, the improvement in accuracy will be better and much clear.

Table-2
The accuracy of the Maximum Likelihood Estimator (MLE) for Gaussian

Data set	Maximum Likelihood Estimator(MLE) Gaussian			
	w=1%	W=5%	W=10%	W=20%
Page block	0.949	0.950	0.950	**0.950**
Pima diabetes	0.767	0.771	0.771	**0.774**
Breast cancer	0.947	0.948	**0.959**	0.951
heart	0.837	0.841	0.841	**0.844**
ionosphere	0.886	**0.898**	0.889	0.883

Table-3
The accuracy of the Maximum Likelihood Estimator (MLE) for uniform

Data set	Maximum Likelihood Estimator(MLE) Uniform			
	w=1%	W=5%	W=10%	W=20%
Page block	0.950	0.950	0.949	**0.951**
Pima diabetes	0.768	0.769	**0.772**	0.769
Breast cancer	0.953	0.952	0.950	**0.963**
heart	0.837	0.837	**0.848**	0.841
ionosphere	0.889	0.892	0.883	**0.892**

Table-4
The accuracy of the probabilities density function (PDF)

Data set	Original (average)	PDF-Gaussian Distribution			
		w=1%	W=5%	W=10%	W=20%
Page block	0.949	0.950	0.950	0.949	0.950
Pima diabetes	0.763	0.766	0.767	0.767	0.775
Breast cancer	0.943	0.949	0.954	0.946	0.959
heart	0.837	0.844	0.837	0.841	0.841
ionosphere	0.878	0.883	0.883	0.889	0.895

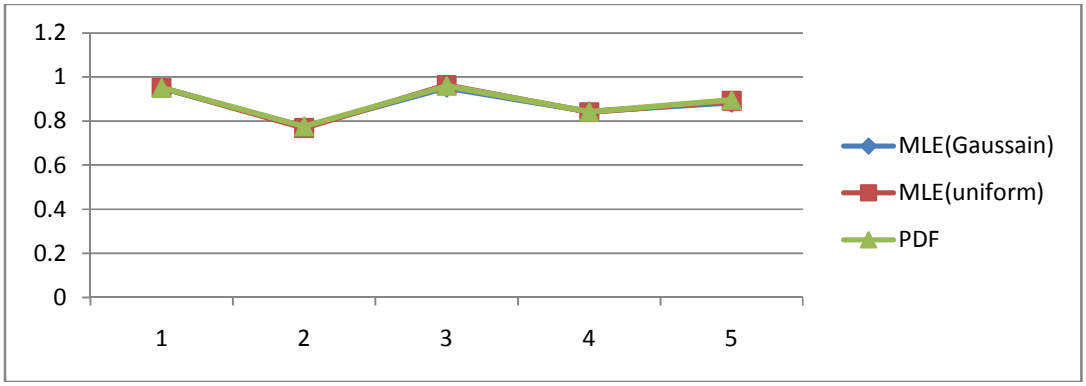

Figure-1
The relationship between MLE and PDF for the data sets

Conclusion

In this study, the standard traditional LR classification algorithm which is constructed to handle the data with single data point is extended to handle data whose objects are numerical with uncertainty. The MLE of the Gaussian and the uniform distributions are computed and used to convert the data to uncertainty. According to the experimental results of the study,

handling data uncertainty using the MLE results in an improvement in the accuracy for both distributions

In the same way, handling data uncertainty with PDF results in significant improvement in accuracy for LR classification. Finally it can be concluded that LR can successfully handle the case when the data is uncertain.

Analyzing uncertainty data can produce classifiers with higher accuracy as compared to using traditional LR algorithm which uses average as a representation of uncertainty, therefore it is recommended that the data should be collected and stored as uncertainty data.

Acknowledgements

This work was supported by a grant from Hebei University, Baoding, Hebei, P.R.China. I like to thank the PhD students of the departments of computer Sciences and mathematics for their encouragement, useful discussions, and interest.
Note: this work is completed in Hebei University during My PhD study period.

References

1. Hosmer D.W. and Lemeshow S., Applied logistic regression, 2nd edn. Wiley series in probability and statistics, Wiley, Inc, New York, **(2000)**

2. Menard S., Applied logistic regression analysis, 2nd edn.Sage publications Inc, (**2002)**

3. Neter J., Kutner M.H., Nachtsheim C.J. and Wasserman W., Applied linear statistical models, 4th edn. Irwin, Chicago**, (1996)**

4. Thomas P. Ryan, Modern Regression Methods. 2nd edn. Wiley-Inter science New York, NY, USA, **(2008)**

5. Brzezinski J.R. and Knafl G.J, Logistic regression modeling for context-based classification. Database and Expert Systems Applications, 1999. Proceedings. Tenth International Workshop on, 755-759, 1999doi: 10.1109/DEXA. 1999.795279,(**1999**)

6. Musa A.B., Comparative study on classification performance between support vector machine and logistic regression*, Int J Mach Learn Cybern,* **4(1),** 13-24 (**2013**)

7. Aggarwal C.C., On Density Based Transforms for uncertain Data Mining. In ICDE Conference Proceedings, **(2007)**

8. Cormode G. and McGregor A., Approximation algorithms for clustering uncertain data, In Principle of Data base System (PODS), M. Lenzerini and D. Lembo, Eds. ACM, 191–200 **(2008)**

9. Singh S., Mayfield C., Prabhakar S., Shah R., and Hambrusc S., Indexing categorical data with uncertainty, In ICDE, 616-625, (**2007**)

10. J. Ge, Y. Xia and C. Nadungodage, UNN: A neural network for uncertain data classication," in PAKDD, 449-460 **(2010)**

11. C.C. Aggarwal, A Survey of Uncertain Data Algorithms and Applications. In IEEE Transactions on Knowledge and Data Engineering, **21(5), (2009)**

12. C.C. Aggarwal, On Density Based Transforms for uncertain Data Mining. In ICDE Conference Proceedings, **(2007)**

13. Tsang S., Kao B., Yip K., Ho W. and Lee S, Decision trees for uncertain data. In: International Conference on Data Engineering (ICDE), (**2009**)

14. J. Bi and T. Zhang, Support vector classication with input data uncertainty," inAdvances in Neural Information Processing Systems (NIPS), 161-168 **(2004)**

15. B. Qin, Y. Xia, and F. Li, DTU, A decision tree for uncertain data," in PAKDD, 4-15 **(2009)**

16. Jiangtao Ren, Sau Dan Lee, Xianlu Chen, Ben Kao, Reynold Cheng and David Cheung, Naive Bayes Classification of Uncertain Data, Ninth IEEE International Conference on Data Mining, (**2009**)

17. B. Qin, Y. Xia and F. Li, A Bayesian classier for uncertain data, in ACM Symposium on Applied Computing, 1010-1014 (**2010**)

18. Eliason S., Maximum Likelihood Estimation: Model and Practice **(1993)**

19. Mood, Graybill, Introduction to the Theory of Statistics, 3rd edn. McGraw Hill, New York, USA, 271–358, **(1974)**

20. Silverman B.W., Density estimation for statistics and data analysis, Lodon; chapman and hall **(1986)**

21. Kim S.J., Koh K., Lustig M., Boyd S. and Gorinevsky D., An interior-point method for largescale l1-regularized least squares. IEEE Journal on Selected Topics in Signal Processing, **1(4),** 606–617 **(2007)**

22. Musa A.B., Comparison of ℓ1-regularizion, PCA, KPCA and ICA for Dimensionality Reduction in Logistic Regression*, Int J Mach Learn Cybern.* doi: 10.1007_s13042-013-0171-7, **(2013)**

23. E.L. Lehmann George Casella, Theory of Point Estimation, Second Edition, Springer, Springer-Verlag New York, Inc, 83-114, **(1998)**

24. Koh K., Kim S.J., Boyd S., l1_logreg: A large-scale solver for l1-regularized logistic regression problems. 0.8.2 Available at http://www.stanford.edu/*boyd/l1_logreg/, **(2009)**