e-ISSN: 2349-9745
p-ISSN: 2393-8161
Scientific Journal Impact Fact

International Journal of Modern Trends in Engineering and Research

www.ijmter.com

Evaluation of Discriminative Least Squares Regression and Extreme Learning Machine for multiclass classification

Shrish Dixit
Teachers Colony, Vidisha, MP

Abstract-This paper presents a short survey of multiclass classification techniques and additionally its applications for. After This I make comparison some best multiclass classification methods like Discriminative Least Square Regression for Multiclass classification of multi feature data and Extreme Learning Machine for Regression andMulticlass Classification in which Neural Network is primarily used. The core plan is to review of all multiclass classification method is to find a systematic overview of all methods with their advantages and disadvantages. Lastly we compare two methods for multiclass classification method on various dataset. There are various additional things are added up to Discriminative least square method like concept of ∈ dragging and uses of hadamart product for Regression and classification. With its compact form, this model are often naturally extended for feature selection. On the other hand neural network based ELM provides a unified learning platform with a widespreadtype of feature mappings and can be applied in regression andmulticlass classification applications directly. So, the aim of this paper is to compare these two modified methods for some multiclass datasets.

Index Terms— Discriminative least squares regression, Feature selection, least squares regression, multiclass classification.

I.INTRODUCTION

Least Squares Regression (LSR) could be a widely-used applied mathematics analysis technique. it\'s been custom-made to several real-world things. LSR earns its place as a basic tool owing to its effectiveness for knowledge analysis still as its completeness in statistics theory. Several variants are developed, as well as weighted LSR [1], partial LSR [2], and different extensions (for example, ridge regression [3]). additionally, the utility of LSR has been incontestable in several machine learning issues, like discriminative learning, manifold learning, clustering, semi-supervised learning, multitask learning, multiview learning, multilabel classification, and so on.

The scope of this paper is to demonstrate multi classification task by method of least squares Regression method and Extreme Learning Machinein transient. Here we tend to compare results of each strategies for classification of various multiclass multi feature datasets.

The method of statistical procedure could be a commonplace approach to the approximate resolution of over determined systems, i.e., sets of equations during which there square measure a lot of equations than unknowns. "Least squares" implies that the general resolution minimizes the add of the squares of the errors created within the results of each single equation. The foremost necessary application is in knowledge fitting. The simplest slot in the least-squares sense minimizes the add of square residuals, a residual being the distinction between associate degree ascertained price and also the fitted price provided by a model. once the matter has substantial uncertainties within the variable (the \'x\' variable), then statistical procedure and statistical procedure strategies have problems; in such cases, the methodology

@IJMTER-2014, All rights Reserved

needed for fitting errors-in-variables models could also be thought-about rather than that for statistical procedure. Least squares method of statistical procedure issues fall under 2 categories: linear or normal least squares and non-linear least squares, betting on whether or not or not the residuals square measure linear all told unknowns. The linear least-squares drawback happens in statistical method analysis; it\'s a closed-form resolution. A closed-form resolution (or closed-form expression) is any formula that may be evaluated in a very finite variety of normal operations. The non-linear drawback has no closed-form resolution and is typically solved byunvaried refinement; at every iteration the system is approximated by a linear one, and therefore the core calculation is comparable in each cases.
Linear regression was the first type of regression analysis to be strictly studied. Given a data set {**x***i*} *Ni*=1 ϵR*m*and adestination set $\{\mathbf{y}i\}_{i=1}$R*c*, where **y***i*is the image vector of **x***i*, the popularly-used regularization for linear regression can be addressed as an optimization problem

$$min_{w,b} \sum_{i=1}^{n} ||W^T x_i + b - y_i||_2^2 + \lambda W_F^2 \qquad (1)$$

Where W ϵRm×c and b ϵR*c*are to be estimated and λ is a regularization parameter, $\|\cdot\|_2$ denotes the L2 norm, and $\|\cdot\|_F$stands for the Frobenius norm of matrix.
In knowledge analysis, (1) is usually applied to knowledge fitting wherever every yiis never-ending observation. Once it\'s utilized for knowledge classification, yiis manually assigned as "+1/−1" for two-class issues or a category label vector for multiclass issues. For classification tasks, it's desired that, geometrically, the distances between knowledge points in numerous categories square measure as giant as potential once they are remodeled. The motivation behind this criterion is extremely kind of like those used for distance live learning [4], [5].
In the past 20 years, owing to their stunning classificationcapability, support vector machine (SVM) [11] andits variants [12]–[14] are extensively employed in classificationapplications. SVM has 2 main learning features: 1) In SVM, The coaching information area unit 1st mapped into the next dimensionalfeature area through a nonlinear feature mapping functionφ(x), and 2) the quality improvement technique is then accustomed realize the answer of increasing the separating margin oftwo completely different categories during this feature area whereas minimizing thetraining errors. With the introduction of the epsilon-insensitive loss operate, the support vector technique has been extended tosolve regression issues [15].

Extreme learning machine (ELM) [16]–[20] studies a muchwider sort of "generalized" single-hidden-layer feedforward network (SLFNs) whose hidden layer neednot be tuned. ELM has been attracting the attentions from a lot of and a lot of researchers [21]–[22]. ELM was originally developedfor the single-hidden-layer feedforward neural networks [16]–[20] then extended to the "generalized" SLFNs that maynot be nerve cell alike [17], [18].

$$F(\mathbf{x}) = \mathbf{h}(\mathbf{x})\boldsymbol{\beta} \quad (2)$$

Where h(x) is that the hidden-layer output comparable to theinput sample x and β is that the output weight vector between the hidden layer and therefore the output layer. One among the salient featuresof ELM is that the hidden layer need not be tuned. Basically,ELM originally proposes to use random procedure nodesire the hidden layer, which are freelance of the trainingdata. Completely different from ancient learning algorithms for a neural variety of SLFNs [23], ELM aims to succeed in not solely the littlest coaching error however additionally the littlest norm of output weights.ELM [12], [13] and its variants [14]–[16], [24], [25] mainly focus on the regression applications. Latest development ofELM has shown some relationships between ELM and SVM [21], [22], [26].

II. INTRODUCTION OF REGRESSION ANALYSIS

In statistics, multivariate analysis may be applied math method for estimating the relationships among variables. It includes several techniques for modeling and analyzing many variables, once the main focus

@IJMTER-2014, All rights Reserved

is on the link between a variable and one or a lot of freelance variables. A lot of specifically, multivariate analysis helps one perceive however the everyday worth of the variable (or \'Criterion Variable\') changes once anybody of the freelance variables is varied, whereas the opposite freelance variables square measure control fastened. Most ordinarily, multivariate analysis estimates the conditional expectation of the variable given the freelance variables – that is, the typical worth of the variable once the freelance variables square measure fastened. Less unremarkably, the main focus is on a quantile, or different location parameter of the conditional distribution of the variable given the freelance variables. All told cases, the estimation target may be a operate of the freelance variables referred to as the regression operate. In multivariate analysis, it is additionally of interest to characterize the variation of the variable round the regression operate, which might be represented by a likelihood distribution.

Regression analysis is wide used for prediction and statement, wherever its use has substantial overlap with the sector of machine learning. Multivariate analysis is additionally accustomed perceive that among the independent variables area unit associated with the variable, and to explore the varieties of these relationships.

Regression models involve the subsequent variables:

The unknown parameters, denoted as β, which can represent a scalar or a vector.

The freelance variables X.

The variable, Y.

A regression model relates Y to a perform of X and β. $Y \approx f(X, \beta)$

There are many types of Regression

1. **Statistical regression** - fits linear and nonlinear models with one predictor. Includes each statistical method and resistant strategies.
2. **Box-Cox Transformations** - fits a linear model with one predictor, wherever the Y variable is remodeled to realize approximate normality.
3. **Polynomial Regression** - fits a polynomial model with one predictor.
4. **Activity Models** - fits a linear model with one predictor then solves for X given Y.
4. **Multiple Regressions** - fits linear models with 2 or additional predictors. Includes associate degree possibility for forward or backward stepwise regression and a Box-Cox or Cochrane-Orcutt transformation.
5. **Comparison of Regression Lines** - fits regression lines for one predictor at every level of a second predictor. Tests for vital variations between the intercepts and slopes.
6. **Regression Model choice** - fits all potential regression models for multiple predictor variables and ranks the models by the adjusted R-squared or Mallows\' Cp data point.
7. **Ridge Regression** - fits a multivariate analysis model employing a methodology designed to handle correlative predictor variables.
8. **Nonlinear Regression**- fits a user-specified model involving one or additional predictors.
9. **Partial least squares** - fits a multivariate analysis model employing a method that permits additional predictors than observations.

Out of those models a number of them area unit used for various machine learning issues like multiclass classification, feature choice, depth estimation of second faces pictures and etc.

III. CONSTRAINED-OPTIMIZATION-BASED ELM

ELM [12]–[14] was originally projected for the singlehidden-layer feedforward neural networks and was then extended to the generalized SLFNs wherever the hidden layer neednot be somatic cell alike [18],

@IJMTER-2014, All rights Reserved

[19]. In ELM, the hidden layer need not be tuned. The output perform of ELM for generalizedSLFNs (take one output node case as associate example) is

$$FL(\mathbf{x}) = \sum_{i=1}^{l} \beta i h i(\mathbf{x}) \;=\; \mathbf{h}(\mathbf{x})\boldsymbol{\beta} \qquad (17)$$

Where β = [β1, . . . , βL]T is that the vector of the output weights between the hidden layer of L nodes and also the output node andh(x) = [h1(x), . . . , hL(x)] is that the output (row) vector of the hidden layer with relevancy the input x. h(x) truly maps thedata from the d-dimensional input area to the L-dimensionalhidden-layer feature area (ELM feature space) H, and thus, h(x) is so a feature mapping. For the binary classificationapplications, the choice operate of

ELM is$fL(\mathbf{x})$ = sign ($\mathbf{h}(\mathbf{x})\,\boldsymbol{\beta}$). (18)

Different from ancient learning algorithms [23], ELM tends to achieve not solely the littlest coaching error however also the smallest norm of output weights. In line with Bartlett's theory [26], for feedforward neural networks reaching smaller training error, the smaller the norms of weights ar, the better generalization performance the networks tend to possess. We conjecture that this could be faithful the generalized SLFNs where the hidden layer might not be nerve cell alike [18], [19]. ELM is to minimize the coaching error further because the norm of the output weights [15], [16]

Minimize: $\|\mathbf{H\beta} - \mathbf{T}\|^2$ **and** $\|\boldsymbol{\beta}\|$ (19)

Where **H** is the hidden-layer output matrix

$$H = \begin{bmatrix} h(x1) \\ \vdots \\ h1(X_N) \end{bmatrix} = \begin{bmatrix} h(x1) & \dots & h1(x1) \\ \vdots & & \\ \vdots & \vdots & \vdots \\ h1(X_N) & \vdots & h1(X_N) \end{bmatrix}$$

To minimize the norm of the output weights $\|\boldsymbol{\beta}\|$ is actually to maximize the distance of the separatingmargins of the two different classes in the ELM feature space:

$2/\|\boldsymbol{\beta}\|$.

The minimal norm least square method instead of the standardoptimization method was used in the original implementationof ELM [15], [16]

$$\boldsymbol{\beta} = \mathbf{H}\dagger\mathbf{T} \qquad (21)$$

Where H† is that the Moore–Penrose generalized inverse of matrixH [28], [29]. Totally different strategies are often accustomed calculate the Moore–Penrose generalized inverse of a matrix: orthogonal projection technique, orthogonalization technique, repetitious technique, and singular price decomposition (SVD) [29]. The orthogonal projection technique [29] are often employed in 2 cases: once HTH is nonsingular and H† = (HTH) −1HT, or once HHT is nonsingular and H† = HT (HHT) −1.

According to the ridge regression theory [30], one will add a positive price to the diagonal of HTH or HHT; the resultant answer is stable and tends to own higher generalization performance. Toh and Deng et al. [21] have studied the performance of ELM with this sweetening below the sigmoid additive style of SLFNs. This section extends such study to generalized SLFNs with a distinct style of hidden nodes (feature mappings) still as kernels.

There is a niche between ELM and LS-SVM/PSVM, and it is not clear whether or not there\'s some relationship between ELM and LS-SVM/PSVM. This section aims to fill the gap and build the relationship between ELM and LS-SVM/PSVM.

@IJMTER-2014, All rights Reserved

IV. PERFORMANCE VERIFICATION

This section compares the performance of various algorithms (Discriminative statistical method Regression and ELM) in real-world benchmark regression, binary, and multiclass classification information sets. So as to check the performance of the projected ELM with varied feature mappings in super tiny information sets.

A. Benchmark Data Sets

In order to extensively verify the performance of different algorithms, wide forms of knowledge sets are tested in our simulations, that area unit of little sizes, low dimensions, large sizes, and/or high dimensions. These knowledge sets embody twelve binary classification cases, twelve multiclassification cases, and twelve regression cases. Most of the info sets area unit taken from UCI Machine Learning Repository [31] and Statlib [32].

1) Binary class data Sets: The twelve binary class data sets (cf. Table II) will be classified into four teams of data:1) data sets with comparatively tiny size and low dimensions,e.g., Pima Indians diabetes, Statlog Australian credit,BupaLiver disorders [31], and Banana [33];

2) Data sets with comparatively tiny size and high dimensions, e.g., leucaemia data set [34] and colon microarray information set[35];

3) Data sets with comparatively giant size and low dimensions, e.g., Star/Galaxy-Bright data set [35], Galaxy Dim dataset [35], and mushroom data set [31];

4) Data sets with giant size and high dimensions, e.g., adult data set [31].

2) Multiclass data Sets: The twelve multiclass data sets (cf.Table III) will be classified into four groups of data as well:

1) Data sets with comparatively tiny size and low dimensions, e.g., Iris, Glass Identification, and Wine [31];

2) Data sets with comparatively medium size and medium dimensions, e.g., Vowel Recognition, StatlogVehicleSilhouettes,andStatlog Image Segmentation [31];

3) Data sets with comparatively giant size and medium dimensions, e.g., letter and shuttle [31];

4) Data sets with large size and/or large dimensions, e.g.,DNA, Satimage[31], and US Postal Service [34].

TABLE II-Specification of Multiclass Classification Problems

Data set	Total num.	Train. num.	Classes	Features
Vehicle	846	340	4	18
AT&T	400	160	40	644
AR	840	360	120	768
Usps	2000	800	10	256
Glass	214	142	9	6
Wine	178	118	13	3
Iris	150	100	4	3

B. Simulation Environment Settings

The simulations of various algorithms on all the info sets except for Adult, Letter, Shuttle, and independent agency data sets area unit meted out in MATLAB seven.0.1 surroundings running in Core a pair of Quad, 2.66-GHZ central processing unit with 2-GB RAM. The codes used for SVM and LS-SVM area unit downloaded from [55] and [56], severally.

Simulations on massive information sets (e.g., Adult, Letter, Shuttle, and USPS data sets) area unit carried out during a superior computer with a pair of.52-GHz central processing unit and 48-GB RAM.

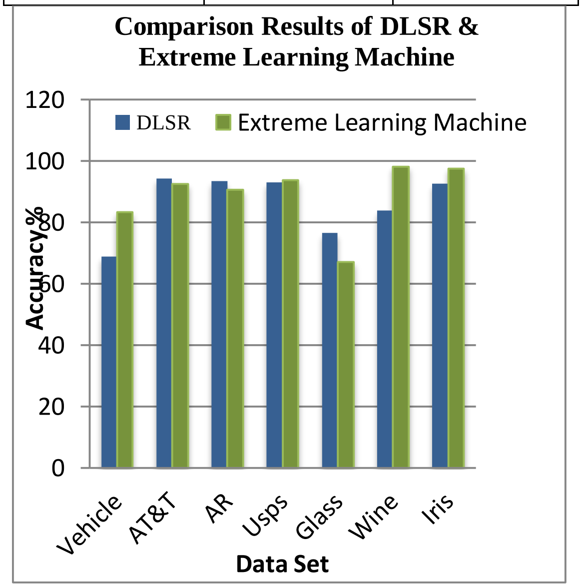
@IJMTER-2014, All rights Reserved

V. RESULT ANALYSIS

We have collected different multiclass dataset as defined in table II. Then we perform experiment i.e., classification of these dataset using method Least Square Regression and Extreme Learning machine, and the results of experiment are shown in table III.

TABLE III-Comparison of Accuracy of Discriminative Least Square Regression and Extreme Learning Machine

Data set	**DLSR**	**Extreme Learning Machine**
Vehicle	68.9302	**83.48**
AT&T	**94.4167**	92.56
AR	**93.4271**	90.77
Usps	93.0667	**93.81**
Glass	**76.56**	67.12
Wine	83.98	**98.17**
Iris	92.67	**97.6**



Graph I Comparison of results

In graph I we compare results of Discriminate Least Square regression and extreme learning machine classifier.

VI. CONCLUSION

In this paper we have a tendency to make a comparison of different multiclass classifier. We have performed our experiments on different datasets. Next we have a tendency to consider least sq. Regression technique and its application space. We have a tendency to brief a typical steps and technique to realize high accuracy as an example we have a tendency to introduce ideas of ϵ dragging in multiclass classification to separate several classes with efficiency. Next as we compare our results with Extreme learning machine, so our tendency is to discover a best classifier algorithm for multiclass classification problem. From the results as shown in Table III and graph I, it is clear that in some of dataset where no. of features are more Least Square regression having better results, on the other hand if number of feature is less extreme learning machine is better idea for multiclass classification.

@IJMTER-2014, All rights Reserved

REFRENCES

1. T. Strutz, Data Fitting and Uncertainty: A Practical Introduction to Weighted Least Squares and Beyond. Wiesbaden, Germany: Vieweg, 2010.
2. S. Wold, H. Ruhe, H. Wold, and W. Dunn, III, “The collinearity problem in linear regression, the partial least squares (PLS) approach to generalized inverse,” J. Sci. Stat. Comput., vol. 5, no. 3, pp. 735–743, 1984.
3. N. Cristianini and J. Shawe-Taylor, an Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, U.K.: Cambridge Univ. Press, 2000.
4. K. Weinberger, J. Blitzer, and L. Saul, “Distance metric learning for large margin nearest neighbor classification,” in Advances in Neural Information Processing Systems 18. Cambridge, MA: MIT Press, 2006, pp. 1473–1480.
5. P. Schneider, K. Bunte, H. Stiekema, B. Hammer, T. Villmann, and M. Biehl, “Regularization in matrix relevance learning,” IEEE Trans. Neural Netw., vol. 21, no. 5, pp. 831–840, May 2010.
6. J. Leski, “Ho–Kashyap classifier with generalization control,” Pattern Recognit. Lett., vol. 24, no. 14, pp. 2281–2290, 2003.
7. R. Duda, P. Hart, and D. Stork, Pattern Classification, 2nd Ed. New York: Wiley, 2001.
V. N. Vapnik, the Nature of Statistical Learning Theory, 2nd Ed. NewYork: Springer-Verlag, 2000.
8. Guo-Zheng Li; Xue-Qiang Zeng, Jack Y. Yang Mary Qu Yang, “Partial Least Squares Based Dimension Reduction with Gene Selection for Tumor Classification”.
9. XinyuanCai, Chunheng Wang, Baihua Xiao, Xue Chen, Ji Zhou, “Regularized Latent Least Square Regression for Cross Pose Face Recognition”, Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence.
10. C. Cortes and V. Vapnik, “Support vector networks,” Mach. Learn.,vol. 20, no. 3, pp. 273–297, 1995.
11. J. A. K. Suykens and J. Vandewalle, “Least squares support vectormachine classifiers,” Neural Process. Lett., vol. 9, no. 3, pp. 293–300,Jun. 1999.
12. G. Fung and O. L. Mangasarian, “Proximal support vector machine classifiers,”inProc. Int. Conf. Knowl. Discov. Data Mining, San Francisco,CA, 2001, pp. 77–86.
13. Y.-J. Lee and O. L. Mangasarian, “RSVM: Reduced support vectormachines,” in Proc. SIAM Int. Conf. Data Mining, Chicago, IL, Apr. 5–7,2001.
14. H. Drucker, C. J. Burges, L. Kaufman, A. Smola, and V. Vapnik, “Support vector regression machines,” in Neural Information Processing Systems 9, M. Mozer, J. Jordan, and T. Petscbe, Eds. Cambridge, MA: MIT Press, 1997, pp. 155–161.
15. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Anew learning scheme of feedforward neural networks,” in Proc. IJCNN,Budapest, Hungary, Jul. 25–29, 2004, vol. 2, pp. 985–990.
16. G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” Neurocomputing, vol. 70, no. 1–3, pp. 489–501, Dec. 2006.
17. G.-B. Huang, L. Chen, and C.-K. Siew, “Universal approximation usingincremental constructive feedforward networks with random hiddennodes,” IEEE Trans. Neural Netw., vol. 17, no. 4, pp. 879–892, Jul. 2006.
18. G.-B. Huang and L. Chen, “Convex incremental extreme learning machine,”Neurocomputing, vol. 70, no. 16–18, pp. 3056–3062, Oct. 2007.
19. G.-B. Huang and L. Chen, “Enhanced random search based incremental extreme learning machine,” Neurocomputing, vol. 71, no. 16–18,pp. 3460–3468, Oct. 2008.
20. X. Tang and M. Han, “Partial Lanczos extreme learning machine forsingle-output regression problems,” Neurocomputing, vol. 72, no. 13–15,pp. 3066–3076, Aug. 2009.
21. Q. Liu, Q. He, and Z. Shi, “Extreme support vector machine classifier,”Lecture Notes in Computer Science, vol. 5012, pp. 222–233, 2008.
22. B. Frénay and M. Verleysen, “Using SVMs with randomised featurespaces: An extreme learning approach,” in Proc. 18th ESANN, Bruges,Belgium, Apr. 28–30, 2010, pp. 315–320.
23. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representationsby back-propagation errors,” Nature, vol. 323, pp. 533–536, 1986.
24. G.-B. Huang, Q.-Y. Zhu, K. Z. Mao, C.-K. Siew, P. Saratchandran, andN. Sundararajan, “Can threshold networks be trained directly?” IEEETrans. Circuits Syst. II, Exp. Briefs, vol. 53, no. 3, pp. 187–191,Mar. 2006.
25. N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, “Afast and accurate on-line sequential learning algorithm for feedforwardnetworks,” IEEE Trans. Neural Netw., vol. 17, no. 6, pp. 1411–1423,Nov. 2006.
26. G.-B. Huang, X. Ding, and H. Zhou, “Optimization method based extremelearning machine for classification,” Neurocomputing, vol. 74, no. 1–3,pp. 155–163, Dec. 2010.
27. P. L. Bartlett, “The sample complexity of pattern classification with neuralnetworks: The size of the weights is more important than the size of thenetwork,” IEEE Trans. Inf. Theory, vol. 44, no. 2, pp. 525–536, Mar. 1998.
28. D. Serre, Matrices: Theory and Applications. New York: Springer-Verlag, 2002.
29. C. R. Rao and S. K. Mitra, Generalized Inverse of Matrices and ItsApplications. New York: Wiley, 1971.

30. A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimationfor non-orthogonal problems," Technometrics, vol. 12, no. 1, pp. 55–67,Feb. 1970.
31. C. L. Blake and C. J. Merz, "UCI Repository of Machine LearningDatabases," Dept. Inf. Comput. Sci., Univ. California, Irvine, CA, 1998.[Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html
32. M. Mike, "Statistical Datasets," Dept. Statist. Univ. Carnegie Mellon, Pittsburgh, PA, 1989. [Online]. Available: http://lib.stat.cmu.edu/datasets/
33. A. Bordes, S. Ertekin, J. Weston, and L. Bottou, "Fast kernel classifierswith online and active learning," J. Mach. Learn. Res., vol. 6, pp. 1579–1619, Sep. 2005.
34. J. J. Hull, "A database for handwritten text recognition research," IEEETrans. Pattern Anal. Mach. Intell., vol. 16, no. 5, pp. 550–554, May 1994.
35. J. Li and H. Liu, "Kent Ridge Bio-Medical Data Set Repository," SchoolComput. Eng., Nanyang Technol. Univ., Singapore, 2004. [Online]. Available:http://levis.tongji.edu.cn/gzli/data/mirror-kentridge.html.

@IJMTER-2014, All rights Reserved