Distributed Linear Programming for Weblog Data using Mining Techniques in Distributed Environment

K. Suresh
MCA (M.Phil)
M.Phil Research Scholar
Department of Computer Applications
Karunya University
Coimbatore, India

Dr Sujni Paul
MCA,M.Phil,Ph.D
Associate professor
Department of Computer Applications
Karunya University
Coimbatore, India

ABSTRACT

Distributed learning discusses different strategies in which learners can communicate with each other. The different strategies are data analysis, predicting future learner in an efficient way to access the learning methods. In this paper the distributed learning has proposed an optimized solution for the fore coming learners. The idea of distributed learning is to analyse the weblog data traversed by the previous learners. Data mining is a process of mining the previously unknown data to make shape up useful knowledge or patterns from large databases. The distributed linear programming is the mathematical approach used in this paper to classify web sites from the weblog data for a specific purpose. Distributed learning approach is used in support vector machine and linear regression method. This paper recommends the fore coming learners to go through the identified web links in order to get high score.

Keywords

E-learning, weblog data, web mining, linear regression, distributed mining.

1. INTRODUCTION

In this paper it explores the collection of web log data, analysis, and visualization of complex data which plays major role in Data mining research and business. Powerful tools obtained from applied statistics, mathematics, and computational methods are used to uncover the meaning behind complex data sets [1][2]. The Data Mining and Information analysis are integrates these from linear programming concepts to provide web designer with statistical information and theoretical basis knowledge for approaching challenging data analysis problems [4]. Web designer trace and learn how to develop and observing analysis from making predictions, to search through large collections of data for rare and unexpected patterns.

The student's details are retrieved from the weblog and mining process is done in distributed environment. Consider an on-line test for students and this test are conducted by some institutions which may consist of N-number of colleges. From this test result the institutions can get the entire toppers list from each and every college to make the other students use the same sites to get good score [11][12]. Now the concern top scorer student in college have to retrieve the web log data for make a prediction about the top scorers and this will be used for the fore coming batch of students. This information mining can be done through the Linear Regression and support vector machine. Mining result will be useful for the students appearing for the fore coming online test. From this online test they can easily predict which source of website is very useful for the online test and also omit which website is given only limited information.

The objective of this paper is Mathematical Linear programming approaches to the fundamental problem like feature selection (i.e.) which web site is efficient for e-learning in cluster of website. The feature selection problem considered is that of discriminating between two methods while recognizing irrelevant and not recommended features. This creates effective model that often generalizes better to new unseen data [3]. This new distributed e-learning discuss about previous learners web log data who are all top scorer in college. So retrieve top scorer weblog data for mining. From this extracted data can recommend the fore coming learners and also for present learners through weblog data [13][14]. A mathematical linear programming formulation of this problem is proposed that is mathematically justifiable and computationally implementable in a finite number of steps. A resulting simplex Algorithm is utilized to discover very useful survival.

2. RELATED WORKS

The analysis of Web log which talks about to advices website owner about a better way to improve the offer, information about what user has faced whether it is problems occurred to the users, and even about problems for the security of the site. Traces about hacker attacks or heavy use in particular intervals of time may be really useful to configure the server and adjust the Web site from the analysis of weblog data [7]. Parallel and distributed computing which talks about expected to reduce current mining methods from the sequential method. Parallel and distributed method can provide the massive datasets, and improving the response time. The main challenges include synchronization and communication minimization, work-load balancing, finding good data layout and data decomposition, and disk I/O minimization, which is especially important for data mining [5]. Distributed data mining and agents which talks about multi agent system, the major drawback in single system is privacy, limited distributed nodes and bandwidth. In this work it broadly discusses about the connection between Distributed Data Mining and Multi Agent System. It focuses on distributed clustering algorithms and it's developed into applications in multi-agent-based problem solving method. This paper discusses one application domain, sensor networks, and potential shortcomings of the current algorithms. This work presented privacy and presents a new algorithm for privacy-preserving clustering [18]. A load-balanced distributed parallel mining algorithm which talks about Parallel and distributed Apriori algorithm. The older Apriori method takes long time to find the frequent patterns when the database contains a large number of transactions. It's also applicable for parallel and distributed techniques to effectively speed-up the mining process. Its goal is to reduce the frequency of database scans and to balance the computation loads among participated computing nodes. In the proposed method, a database has only to be scanned once because metadata are stored in the form of Transaction Identifiers (TIDs). This approach takes item set counts into consideration to improve load balancing as well as to reduce idle time of processors. In this proposed system algorithm, each transaction has a unique Transaction Identifier, called TID. By using hash functions to store TIDs in a table structure, the number of item sets can be quickly calculated without the need of re-scanning the database [19].

3. SUPPORT VECTOR MACHINE

A support vector machine is (SVM) is a classification for different data to build a hyperplane, this new constructed method can separates the data into two different categories. This classification is identified error-bound analysis has been motivated for support vector data. Support vector machines had significant success in numerous learning tasks. The most machine learning algorithms, they are generally applied using a randomly selected training set classified in advance [4][18]. This method is defined over a vector space in which the problem is to find a decision surface that best separates the data vectors into two classes. In this simplest linear form, an SVM is a hyper plane that separates a set of positive examples from a set of negative examples with maximum margin. By using hyperplane it will classify into linear attribute and non-linear attribute data.

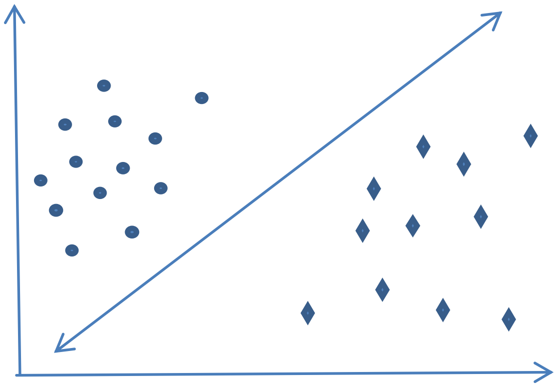


Figure.1 Support vector machine

In SVM operator, a predictor variable (i. e) unknown data is known as an attribute, and a transformed attribute that is used to define the hyper plane is called a feature. The task of choosing the most suitable representation is known as feature selection of data to show what type of it. A set of features that describes one case is called a vector. So the goal of SVM modelling is to find the optimal hyper plane that separates clusters of vector with one category of the target variable are on one side of the plane and cases with the other category are on the other size of the plane. The vectors near the hyperplane are the support vectors.
For example the above figure 1 represents the sample diagram for Support Vector Machine. The figure has two different categorized data (i.e.) oval shape represents Number of Hits and Diamond shape represents Duration of time used by students. These two different data are clearly differentiated by the hyper plane line. It is represented between the two data. From this diagram the support vector data are identified, which is the some data are placed near to the hyper lane and this data is known as the support vector.

In this work, the SVM is used to classify the student's weblog data from one student to another student and it also compares the student outcome from one location to another location. From the classification error bound analysis is made by the support vector data. Support vector machines had significant success in numerous learning tasks. This information mining can be done through the Linear Regression and support vector machine. Mining result will be useful for the students appearing for the fore coming online test. From this online test they can easily predict which source of website is very useful for the online test and also omit which website is given only limited information.

4. SIMPLE LINEAR REGRESSION ANALYSIS

Regression analysis is a statistical technique that attempts to explore and model the relationship between two or more variables. For example a Website owner wants to know if there is relationship between Visit duration and number of Hits in the website then the Regression analysis forms an important part of the statistical analysis of the data obtained from weblog data. As website owner being the process of optimizing their site, they need the ability to accurately measure the results of your efforts [21][23]. If the website designer continuously make changes before or cannot analyze web status they will not be able to effectively track the website. This analyzing work will be effective if it's done in Linear Regression. The Analyzing attribute are Visit duration, key phrases, Page tracking, web status error report, Number of Hits, How long person using the website.

From the below analysing attribute Number of Hits and duration hour to read the particular website are going to use in this Linear Regression Analysis. A linear regression model attempts to explain the relationship between two or more variables using a straight line.

Table 1. Weblog Data Used in regression Calculation

S.No	**Number of Hits(X)**	**Duration In Minutes(Y)**	**X * Y**	
1	10	100	1000	144
2	15	88	1320	49
3	25	95	2375	9
4	20	80	1600	4
5	35	105	3780	169
6	05	70	350	289
7	20	54	1080	4
8	10	42	420	144
9	33	149	4917	121
10	30	161	4830	64
11	22	67	1474	0
12	24	140	3360	4

13	28	110	3080	36
14	30	125	750	64
15	33	105	3465	121
16	13	122	1586	81
17	19	90	1710	9
18	25	93	2325	9
19	27	188	5076	25
20	29	200	5800	49

5. REGRESSION LINE

The true regression line corresponding to Eqn. (1) is usually never known. However, the regression line can be estimated by estimating the coefficients and $\hat{\beta}_0$ for an observed data set [24] [25]. The estimates, $\hat{\beta}_1$ and $\hat{\beta}_0$, are calculated using least squares. The estimated regression line, obtained using the values of $\hat{\beta}_1$ and β0, is called the fitted line. The least square estimates, $\hat{\beta}_1$ and $\hat{\beta}_0$

$$\hat{\beta}= \sum_{i=1}^{n} y_i\, x_i - \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i)}{(\sum_{i=1}^{n} x_i - \bar{x})^2}$$

→ Equation -1

The least square estimates of the regression coefficients can be obtained for the data in Table

$$\hat{\beta} \ = 50298 - \frac{\frac{(453)(2184)}{20}}{1395}$$

$$\hat{\beta}_1 = 0.68\ 4$$

$$\hat{\beta}_{0\,=}\bar{y} - \quad \hat{\beta}_1\bar{x}$$

$$\hat{\beta}_{0\,=}109 - 0.6 * 22$$

$$\hat{\beta}_{0\,=}95.8$$

Knowing and $\hat{\beta}_0$ the fitted regression line is 95.8

Distributed Problem solve by using simplex Method

10 X1 + 100 X2 + 1000 X3 = 144
20 X1 + 29 X2 + 200 X3 = 5800
12 X1 + 24 X2 + 140 X3 = 3360

These above Equations are fetched from Table – 1 randomly and these values are containing the attribute maximum number of hits and maximum number of duration used by the students for online cat exams [26]. But these are perfectly not applicable for solving in simplex method so according to maximum number of value framing general equations for solving minimization iteration in simplex method.

Z = 8x1 + 10x2 + 7x3
X1 + 3x2 + 2x3 <= 10
X1 + 5x2 + x3 <= 8
X1, x2, x3 >= 0

Table 2. Simplex Iteration - I

1	**3**	**2**	**1**	**0**	**0**	**10**
1	**5**	**1**	**0**	**1**	**0**	**18**
-8	**-10**	**-7**	**0**	**0**	**1**	**0**

Table 3. Simplex Iteration – II

2/5	**0**	**7/5**	**1**	**-3/5**	**0**	**26/5**
1/5	**1**	**1/5**	**0**	**1/5**	**0**	**8/5**
-6	**0**	**-5**	**0**	**2**	**1**	**16**

Table 4. Simplex Iteration – III

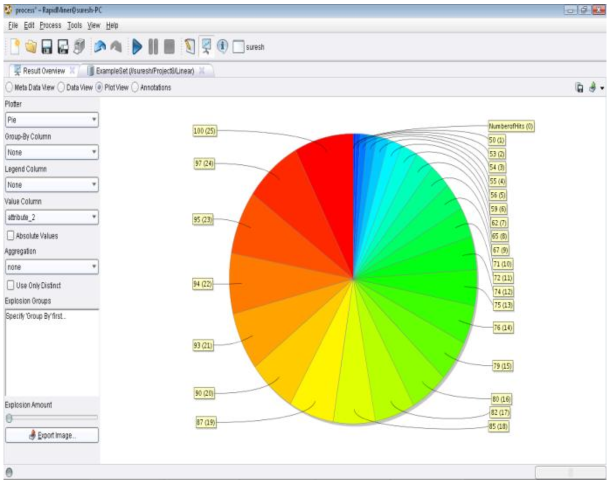
0	**-2**	**1**	**1**	**-1**	**0**	**2**
1	**5**	**1**	**0**	**1**	**0**	**8**
0	**30**	**1**	**0**	**8**	**1**	**64**

X=8, X2=0, X3=0, S1=2, S2=0, Z=64

- From this final simplex tableau, the maximum value of z is 64. Therefore the original solution of minimization problem is z = 64.
- From the above result is derived from three topper students' web log data. In this result web site owner can predict exact web site that's which website students hits most of the time for online exam. Web site owner can also recommended these web sites through online forum of the organization and also not recommended least number of hits in the web site (i.e.) less than of 5 hits**.**

6. RESULTS AND DISCUSSIONS

The frequently accessed websites viewed by top scorer students can be easily retrieved through the mining process. Using data mining tool Tiberius the pictorial representation is shown below. In this tool it mines the weblog data which is fetched from top scorer students. Obtained from this result the extracted information can recommend efficient websites which are related to the online test. This is also useful for current and fore coming learners. By gathering these information student will update the current status of information for the online test.



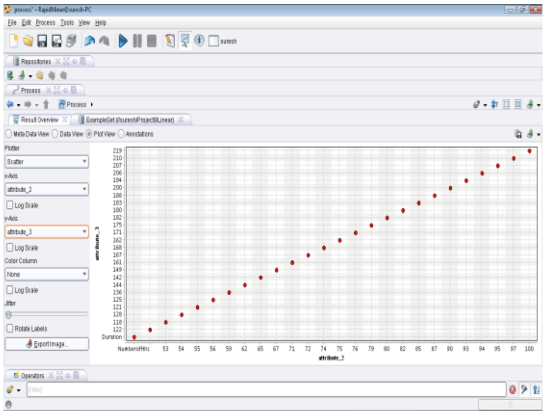
Result Overview | ExampleSet (//suresh/Project8/Linear)

○ Meta Data View ○ Data View ● Plot View ○ Annotations

Plotter: Pie
Group-By Column: None
Legend Column: None
Value Column: attribute_2
□ Absolute Values
Aggregation: none
□ Use Only Distinct
Explosion Groups: Specify 'Group By' first...
Explosion Amount

Figure 2. Pie Chart

From the table-1 web site owner can revealed that who are the students used website for how many minutes. The duration of time can be easily viewed through Pie chart. The pie chart represented maximum duration in large space and minimum duration in small size.



Result Overview | ExampleSet (//suresh/Project8/Linear)

○ Meta Data View ○ Data View ● Plot View ○ Annotations

Plotter: Scatter
x-Axis: attribute_2
□ Log Scale
y-Axis: attribute_3
□ Log Scale
Color Column: None
□ Log Scale
Jitter
□ Rotate Labels

Figure 3. Linear Regression Data Plot by using Values from Table-1

The above figure represented as the linear regression model that can be used to explain the relation between X and Y that is seen on the scatter plot above. In this model, the mean value of Y (abbreviated as E(Y)) is assumed to follow the linear relation β0 + β1x [16][17].

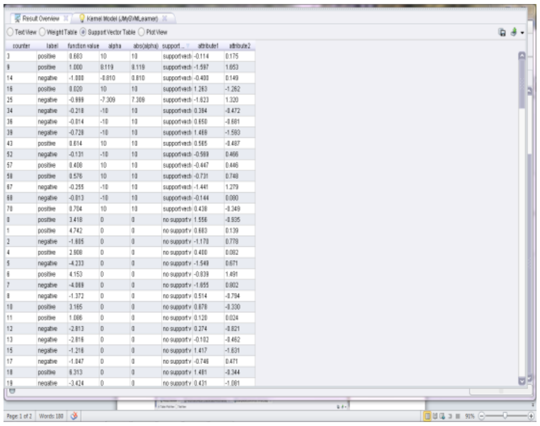
$$E(Y) = \beta 0 + \beta 1x \quad \text{--------}\rightarrow \text{Equation -2}$$

The actual values of Y, (which are observed as yield from the chemical process from time to time and are random in nature), are assumed to be the sum of the mean value, E(Y) E(Y), and a random error term

$$Y = E(Y) + €$$

$$= \beta 0 + \beta 1x + € \quad \text{---}\rightarrow \text{Equation - 3}$$

The regression model here is called a simple linear regression model because there is just one independent variable, X, in the model. In regression models, the independent variables are also referred to as regression or predictor variables. The dependent variable, Y, is also referred to as the response [6]. The slope, β0 and β1, and the intercept, β0and β1 of the line E(Y) = β 0+ β 1are called regression coefficients [8]. The slope, β1, can be interpreted as the change in the mean value of Y for a unit change in X.



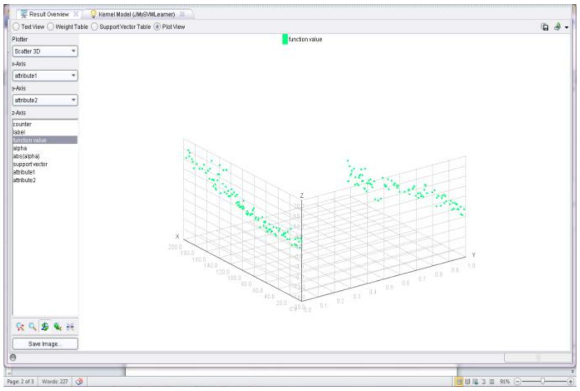
Result Overview | Kernel Model (JMySVMLearner)

○ Text View ○ Weight Table ● Support Vector Table ○ Plot View

counter	label	function value	alpha	abs(alpha)	support ...	attribute1	attribute2
3	positive	0.683	10	10	support vect	-0.114	0.175
9	positive	1.000	8.119	8.119	support vect	-1.597	1.653
14	negative	-1.000	-0.810	0.810	support vect	-0.400	0.149
16	positive	0.020	10	10	support vect	1.263	-1.262
25	negative	-0.999	-7.309	7.309	support vect	-1.623	1.320
34	negative	-0.218	-10	10	support vect	0.394	-0.472
36	negative	-0.014	-10	10	support vect	0.650	-0.661
39	negative	-0.728	-10	10	support vect	1.469	-1.593
43	positive	0.614	10	10	support vect	0.565	-0.487
52	negative	-0.131	-10	10	support vect	-0.569	0.466
57	positive	0.408	10	10	support vect	-0.447	0.446
58	positive	0.576	10	10	support vect	-0.731	0.748
67	negative	-0.255	-10	10	support vect	-1.441	1.279
68	negative	-0.013	-10	10	support vect	-0.144	0.080
70	positive	0.704	10	10	support vect	0.438	-0.349
0	positive	3.418	0	0	no support v	1.556	-0.935
1	positive	4.742	0	0	no support v	0.683	0.139
2	negative	-1.605	0	0	no support v	-1.170	0.778
4	positive	2.908	0	0	no support v	0.400	0.082
5	negative	-4.233	0	0	no support v	-1.549	0.671
6	positive	4.153	0	0	no support v	-0.839	1.491
7	negative	-4.069	0	0	no support v	-1.655	0.802
8	negative	-1.372	0	0	no support v	0.514	-0.794
10	positive	3.165	0	0	no support v	0.878	-0.330
11	positive	1.086	0	0	no support v	0.120	0.024
12	negative	-2.813	0	0	no support v	0.274	-0.821
13	negative	-2.816	0	0	no support v	-0.102	-0.462
15	negative	-1.218	0	0	no support v	1.417	-1.631
17	negative	-1.047	0	0	no support v	-0.746	0.471
18	positive	6.313	0	0	no support v	1.481	-0.344
19	negative	-3.424	0	0	no support v	0.431	-1.081

Figure 4. Kernel mode in support vector machine

Out of 200 it has 15 support vectors. The above diagram represents support vector table by using support vector machine visualization. In this representation the first 15 data which is included in table are considered as a support vector data and the remaining data are categorized into non-support vector data [9]. Because the non-support vector data are not near to the hyper lane vector.



Result Overview | Kernel Model (JMySVMLearner)

○ Text View ○ Weight Table ○ Support Vector Table ● Plot View

Plotter: Scatter 3D
x-Axis: attribute1
y-Axis: attribute2
z-Axis: counter, label, function value, alpha, abs(alpha), support vector, attribute1, attribute2

Figure 5. Kernel mode plotted Function value

The above figure represented the function value for two attribute which is Number of hits and Duration of time spend in web pages during the online exam. The two attributes are clearly plotted in green dots which are shown in the above diagram. So left side of data plotted as number of hits attribute and right side of data was plot as a Duration of Time attribute. According to the support vector these are two attribute categorized into two types. Now the support vector data are calculated according to the hyper lane its will be drawn or mentioned between these two attribute. Implementing an SVM operator with an adequate level of usability and performance result in accuracy is 87.50%

SVM implementation allows data with little data mining expertise to achieve reasonable out-of-the-box results from the 200 vectors. From this the total bias (offset) value is 0.311.

The separate value for each attribute is
W [att1] = 5.352
W [att2] = 5.588

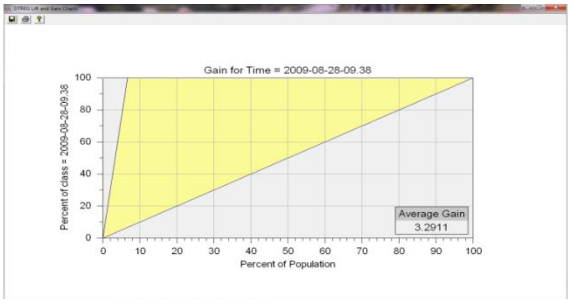


Figure 6. Duration of Time measured in Lift and Gain Method

The above diagram represented lift and gain about weblog data which are two major categorized attributes Number of Hits and Duration of Time. From the table-1 data it represented average gain of time is 3.2911 and this results are predicted according to the time is 8.28.09 so in this time the major number of hits are happened with maximum duration browsed belongs to online exam and if classification is easier in a high-dimensional feature space.

7. REFRENCES

[1] Botia J.A., Garijo, J.R., and Skarmeta.A.f.1998. A Generic Data Mining System:Basic Design And Implementation Guidelines in workshop on Distributed Data mining at the Fourth Intl. Conf. on Data Mining and Knowledge Discovery.

[2] S. Krishnaswamy, S.W. Loke, A. Zaslavsky. Cost Models for Distributed DataMining school of computer science and software engineering , Monash University.

[3] Ning Chen1, Nuno C. Marques1, and Narasimha Bolloju2. AWeb Service-based approach for data mining in distributed environments.

[4] Alex J. Smola and Bernhard Scholkopf .(September 30, 2003). A Tutorial on Support Vector Regression.

[5] Cherkassky and F. Mulier. 1998. Learning from data. john wiley and sons, New York.

[6] Cesar vialardi, Javier Bravo, Leila shafti, Alvaro. Ortigosa. 2009. Recommendation In higher education using data mining techniques.

[7] Agrawal. R., Imielinski, T., and Swami, A. 1993. Mining association rules between sets of items in large database. ACM SIGMOD Conference on Management of data.

[8] Liu M., Wang F.Y., Zeng D., and YangL. 2001. An Overview of world Wide Caching, IEEE International Conference on Systems, Man and Cybematics.

[9] Rousskov, A., and Soloviev, V. 1998. On Performance of Caching Proxies, Short version appears as poster paper in ACM SIGMETRIC'98 Conference.

[10] Liu M., Wang F.Y., Zeng D., and YangL. 2001. An Overview of world Wide Caching, IEEE International Conference on Systems, Man and Cybematics, pp. 3045-3050.

[11] M. Holsheimer and A. Siebes .1994. The search for knowledge in databases. Technical Report CS-R9406, CWI, Netherlands.

[12] R. Kosala and H. Blockeel. June 2000. Web mining research: A survey. SIGKDD Explorations.

[13] B. Masand and M. Spiliopoulou, editors. Advances in Web Usage Mining and User Profiling:Proceedings of the WEBKDD'99 Workshop. Number 1836 in LNAI. Springer Verlag,

[14] O. Zaiane, M. Xin, and J. Han. April 1998. Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs. In Proc. Advances in Digital Libraries Conf. (ADL'98), Melbourne, Australia, pages 1244-158.

[15] H. Mannila and C. Meek. Aug 2000. Global partial orders from sequential data. In Proc. of the 6th Intl. Conf. on Knowledge Discovery and Data Mining (KDD2000), pages 161–168.

[16] R.C Agarwal, C.C Aggarwal and V.V.V. Prasad. A. 2001 tree projection algorithm for generation of frequent item sets. Journal of parallel and distribute computing, 61(3):350-371.

[17] S. Parthasarathy, M. Zaki, and W. Li. August 1998 Memory Placement Techniques for Parallel Association mining in the fourth ACM SIGKDD International Conference on knowledge Discovery on knowledge.

[18] Josenildo C. da Silvaa, Chris Giannellab,, RuchitaBhargavac, HillolKarguptab,d, Matthias Kluscha "Distributed data mining and agents", 2005

[19] Kun-Ming Yu a, Jiayi Zhou b, Tzung-Pei Hong c, Jia-Ling Zhou d A load-balanced distributed parallel mining algorithm, 2009

[20] David meyer, Technishe university at wien, Austria support vector machines Interface to libsvm in package e1071, april 21 2010

[21] C . Cortes and V. Vapnik support vector networks. Machine learning, 1995

[22] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, Neural Networks for Signal Processing VII — Proceedings of the 1997 IEEE Workshop, pages 276 – 285, New York, 1997. IEEE

[23] A. J. Smola. Regression estimation with support vector learning machines Master's thesis, Technische Universit¨at M¨unchen, 1996

[24] M. O. Stitson and J. A. E. Weston. Implementational issues of support vector machines Technical Report CSD-TR-96-18, Computational Intelligence Group, Royal Holloway, University of London, 1996

[25] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation, and signal processing. In M. Mozer, M. Jordan, and T. Petsche, editors, Advances in Neural Information Processing Systems 9, pages 281– 287, Cambridge, MA, 1997. MIT Press.

[26] S. S. Keerthi and C.-J. Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation, 15(7):1667{1689, 2003.

[27] Pearson K. Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. Phil Trans R SocLond Series A 1896; 187:253–318.