Tri Peptide Composition as Effective Classifier of GPCRs with Application of Rough Sets

Kumud Pant
Department of Biotechnology,
Graphic Era University,
Dehradun, India
Pant.kumud@gmail.com

Bhasker Pant
Department of IT,
Graphic Era University,
Dehradun, India
Pantbhaskar2@gmail.com

***Abstract*—GPCRs or G Protein Coupled Receptors are very important molecules and are implicated not only in normal physiological processes but also in diseases. Around 40% of the drugs in the market are targeted towards GPCRs. Through out the world various attempts have been made to analyze and understand these molecules using various attributes and algorithms but large amount of data redundancy and noises in the data set lead either to decrease in classification accuracy as well as increase in dimension which might not provide enough information.**

This is a novel attempt of its kind where rough sets have been used for effectively reducing the dimension of tripeptide composition evaluated on CC Chemokine receptors and their ten (10) different types. The results obtained by implementation of support vector machine (SVM) using software SVMlight has shown more than 5% increase in accuracy for the various subgroups when analysis was performed again with reduct set comprising. The dimension of reducts obtained through Rough sets was only 400 which has helped in removing the redundancy in the data set and also paved a way for analysis of proteomic and genomic data set using attributes with even higher dimensions)

Keywords- G Protein Coupled Receptors (GPCRs), SVMLight, Rough Sets, Dimensionality Reduction

I. INTRODUCTION *(HEADING 1)*

With numerous wet lab experiments conducted through out the world escalation in proteomic and genomic data has caused problem of piling of unannotated data. Hence intervention of other field of studies like computational sciences, mathematical sciences, statistics and their algorithms has come into picture. This interdisciplinary approach has opened up various possibilities where by data generated has been analyzed by using attributes like amino acid composition, dipeptide composition and tripeptide composition. Various algorithms of artificial intelligence like support vector machines neural networks and naïve bayes classifiers have been implemented and used to analyze the data using various attributes. Amino acid composition an attribute which provides a dimension of 20 is the most widely used attribute for protein classification. But it does not consider position effect of various amino acids. Higher dimensional attributes like dipeptide and tripeptide composition on the other hand takes into account the position effect but also face the problem of increase in noise and data redundancy. The attribute of tripeptide composition used for protein analysis has dimension of 8000, incorporates many tripeptides which do not exist in the protein and have value of 0. This attribute can provide us enough information about the kind of amino acid triplets which have the highest tendency to be associated together as well as the positional effect which they exert on the protein but a closer analysis of 8000 dimension reveals that there is a great increase in noises and non informative values in the output. Moreover protein attributes with even higher dimension like tetra peptide composition although provides information regarding propensity of particular tetra peptide to characterize a particular protein class but at the same time adds redundant and un informative content.

Realizing this here we present a novel approach for building rough set rule extractor from tri peptide composition with GPCRs as test data set. The training of the classifier has been done through Support Vector Machine (SVM) which has been implemented through LibSVM and SVMLight. We have chosen SVMs because they are very robust and efficient classifiers which work by building a hyper plane between classes on the basis of attribute chosen for their analysis. SVMs build the hyper plane or decision boundary around the points or support vectors which are the most representative of the class. There are points which lie far apart from the hyper plane which might not be of great use in constructing the decision boundary between classes.

We are taking data set of CC Chemokine GPCRs with 10 groups, for our experimental studies because they are very important molecules which have normal as well as diease implications in organisms. According to an estimate around 40% of the drugs in the market are based on GPCRs and they are the hottest drug targets. CC chemokine receptors (or beta chemokine receptors) are 7 pass transmembrane proteins (7-TM) like other GPCRs that specifically bind and respond to cytokines of the CC chemokine family. To date, ten true members of the CC chemokine receptor subfamily have been described. According to the IUIS/WHO Subcommittee on Chemokine Nomenclature they are named from CCR1 to CCR10. They all work by activating G proteins (1). They have been implicated both directly and indirectly in parasitic infections (2). Similarly they have been found to play role in diseases like obstructive pulmonary disease, kidney diseases to name a few (3, 4). Realizing the importance of these proteins in normal physiological processes and disease state we propose a rough set based most evaluative feature extractor algorithm as an input for Support Vector Machine. The reducts obtained through rough sets are used to classify and predict the above GPCR classes with implementation of SVM through both SVMLight and LibSVM.

Rough sets are based on the assumptions that we can identify any object belonging to a class only through those properties which can be evaluated and redundant properties cannot distinguish two objects. The theory for Rough sets was introduced by Zdzislaw Pawlak in 1982. According to this theory if we can deduce a set of rules for a particular class based on 'Quality of Lower Approximation' and 'Quality of Upper Approximation' then the dependencies of attributes can classify the whole class and removal of redundant values will produce results with same accuracies (6, 7, 9).

Previously tri peptide composition has been used for prediction of Glutathione S-Transferase proteins (5). Since the rough set concept has great use in dimensionality reduction in large databases therefore they have successfully been applied in micro array data for cancer prediction as well (10).

In a novel attempt of its kind we are trying to show the suitability of Rough set estimators in dimensionality reduction with tri peptide composition using dataset of GPCRs. This approach will provide a way to extract the most informative feature and also couple it with informative features from other attributes so that hybrid classifier with optimum set of properties can be built.

II. MATERIALS AND METHODS

A. The Data Set

The dataset for proteins was obtained from SwissProt/UniProt server of Expasy server of Swiss Institute of Bioinformatics. After redundancy removal till 90% and fragment elimination a workable data set of 34 fasta sequences for CCR1, 36 for CCR2, 37 proteins for CCR3, 36 for CCR4, 34 for CCR5, 35 FOR CCR6, 38 for CCR7, 37 proteins for CCR8, 34 proteins for CCR9 and 36 for CCR10. The total data set was of 357 proteins (8).

B. RSES Software

RSES software developed by a team supervised by professor Andrzej Skowron. It is freely available and can be downloaded from http://logic.mimuw.edu.pl/rses. It provides user with the ability for analysis of tabular data sets with use of various methods in particular those based on rough set theory (6)

C. Rough sets

In 1982 Z. Pawlak proposed his algorithm to deal with noises, uncertaininty, fuzzyness and incompleteness in the data set in the form of Rough set theory. In this algorithm with the data set at hand the generalizations can be mined in the form of rules which in turn can be used to validate the generalizations (7, 9).

D. Definition

Let S be an information system formed of 4 elements

$$S = (U, Q, V, f)$$

Where:U - Is a finite set of objects

Q - Is a finite set of attributes

V- Is a finite set of values of the attributes

f - Is the information function so that:

$$f: U \times Q - V.$$

Let P be a subset of Q, $P \subseteq Q$, i.e. a subset of attributes. The indiscernibility relation noted by IND (P) is a relation defined as follows

$$IND (P) = \{< x, y > \in U \times U: f(x, a) = f(y, a), \text{ for all } a \in P\}$$

If $< x, y > \in$ IND (P), then we can say that x and y are indiscernible for the subset of P attributes. U/IND (P) indicate the object sets that are indiscernible for the subset of P attributes.

$$U / IND (P) = \{ U1, U2, \ldots\ldots Um \}$$

Where $Ui \in U$, i = 1 to m is a set of indiscernible objects for the subset of P attributes and $Ui \cap Uj = \Phi$, i, j = 1to m and $i \neq j$. Ui can be also called the equivalency class for the indiscernibility relation. For $X \subseteq U$ and P inferior approximation P1 and superior approximation P1 are defined as follows

$$P1(X) = U\{Y \in U/ IND (P): Y \subseteq Xl\}$$

$$P1(X= U\{Y \in U / INE (P): Y \cap X \neq \Phi \}$$

Rough Set Theory is based on finding reduct from the original set of attributes. This set of reducts represents an equivalent set of characteristics as the original set and can be used with any data mining algorithm . The set of attributes Q from the informational system S = (U, Q, V, f) can be divided into two subsets: C and D, so that $C \subset Q$, $D \subset Q$, $C \cap D = \Phi$. Subset C will contain the attributes of condition, while subset D those of decision. Equivalency classes U/IND(C) and U/IND (D) are called condition classes and decision classes. The degree of dependency of the set of attributes of decision D as compared to the set of attributes of condition C is marked with γc (D) and is defined by

$$\gamma\, c(D)= \frac{|\, POSc(D)|}{|U|}\ , 0 : \gamma c(D) : 1$$

$$|\, POSc(D)|= U \subseteq X$$

$$X \in U/IND\ (D)$$

POSC (D) contains the objects from U that can be classified as belonging to one of the classes of equivalency U/IND (D), using only the attributes in C. If γc (D) = 1 then C determines D functionally. Data set U is called consistent if γc (D) = 1. POSC (D) is called the positive region of decision classes U/IND (D), bearing in mind the attributes of condition from C.

Subset R ⊂ C is a D-reduct of C if POSR (D) = POSC (D) and R has no R' subset, R' ⊂ R so that POSR'. (D) = POSR (D). Namely, a reduct is a minimal set of attributes that maintains the positive region of decision classes U/IND (D) bearing in mind the attributes of condition from C. Each reduct has the property that no attribute can be extracted from it without modifying the relation of indiscernibility. For the set of attributes C there might exist several reducts (7).

E. Discretization

Discretization is the process whereby we convert or partition data which is of continuous nature to a form to nominal form or in intervals. While dealing with rough sets it is very important to discretize the attributes so as to form effective rules. Here we discretize data by generating cuts which in turn reduces the size of the data and allows wide applicability of rules instead of being specific

F. Generation of Reducts

Out of the above data set now the attributes which are common and redundant are removed and only those attributes are reported which can define the entire class in the same way as the entire set of attributes. They are the reducts which represent the minimal set of attributes required to define a class hence a step ahead in reducing noises and bulkiness of the data.

G. Support vector Machines

Support vector machines are a set of supervised learning techniques that can classify data on the basis of construction of hyperplane or sets of hyperplanes in a multi dimensional space. The theory of support vector machine was given by Vapnik et.al. in 1995 (10). In SVMs good separation is achieved by a hyperplane that has the largest distance to the nearest data point of any class also called functional margins. The larger the margin the lower the generalization error of the classifier. SVMs can be applied to both linear and non linear data with adjustments in the kernel parameters. As depicted in Figure 2 below support vectors are the representatives of their class which can be used for building the hyper planes to separate the classes.

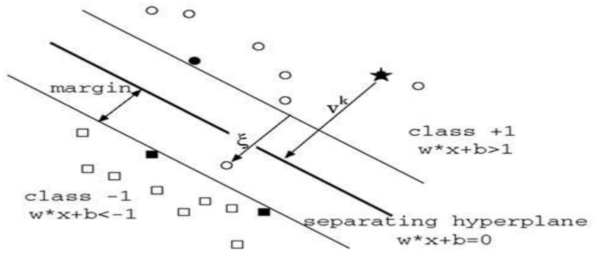

FIGURE 1. DIAGRAM INDICATING SUPPORT VECTORS AS DARKENED POINTS

It is seen that for constructing the hyper planes some data set points might not prove to be very useful. Therefore an optimal classifier can be build by removing these common or redundant attributes

H. A Rough SVM approach for statistical factors for protein classification

Here we present a novel approach whereby we are reporting the utility of rough set theory based dimensionality reduction of tri peptide composition and hereby a decrease in noise. The decreased number of attributes when used for training and testing of the data give comparable and at some places better results when used with SVMs. The results are given below.

III. RESULTS AND DISCUSSION

The optimization algorithms used in SVMlight are described in (Joachims, 2002 and Joachims, 1999) and freely downloadable from http://svmlight.joachims.org/ (13, 14). Implementation of rough set theory in the form of RSES software was done in the form of RSES software available at http://logic.mimuw.edu.pl/~rses/ (6).

When tripeptide composition of GPCRs was trained with LibSVM, five fold cross validation accuracy with value of C and Gamma (default parameters of SVM Radial Basis Function) and the accuracy values with values of C and Gamma are shown in Table1 below.

TABLE 1. RESULTS WITH AND WITHOUT ROUGH SETS

Class (CCR)	*C with TPC*	*Gamma with TPC*	*Accuracy with TPC*	*C with Reducts*	*Gamma with Reducts*	*Accuracy with Reducts*
1	0.03125	0.0078125	97%	8	0.0001220703125	97.5% (Fig. 1)
2	2	0.0078125	95%	8	0.0001220703125	100% (Fig. 2)
3	2	2	96%	8	0.0001220703125	97.5% (Fig. 3)
4	0.03125	0.0078125	95%	8	0.0001220703125	100% (Fig. 4)
5	512	0.0001220703125	95%	8	0.0001220703125	100 (Fig. 5)
6	2	0.0078125	92%	8	0.0001220703125	97.4359% (Fig. 6)
7	512	0.0078125	92.3%	0.03125	0.0078125	97.4359% (Fig. 7)
8	2	2	91%	0.03125	0.0078125	94.8718% (Fig. 8)
9	2	2	91%	2	0.0001220703125	94.8718% (Fig. 9)
10	0.03125	0.0078125	93%	8	0.0001220703125	100% (Fig. 10)

Of the 8000 dimension of TPC after application of Rough Sets only 400 were evaluated to be informative and significant. On manual inspection of the vector comprising 8000 dimension it was found that the composition value of majority of tripeptides was 0. It is especially true in case of smaller proteins like CC Chemokine receptors. In the first protein of CCR2 chemokine with accession ID >GI|3154235 the following composition were found to be empty i.e. with 0 value.

TABLE II. THE TRIPEPTIDES IN DIMENSION 8000 WITH NON ZERO VALUES FOR CCR2 CHEMOKINE ACCESSION ID >GI|3154235

37, 138, 150, 151, 154, 191, 197, 200, 201, 202, 212, 238, 245, 246, 268, 313, 375, 377, 382, 394, 396, 398, 399, 400, 401, 402, 414, 431, 435, 531, 580, 591, 607, 610,667,678, 696, 725, 736, 740, 771, 925, 1011, 1016,1090, 1191, 1213, 1239, 1278, 1331, 1401, 1420, 1477, 1517, 1572, 1591, 1700, 1707, 1732, 1831, 1915, 1917, 1961,1983, 2019, 2034,2136,2166, 2194, 2196, 2216,2232, 2264, 2277, 2411, 2429, 2480,2486, 2562, 2604, 2614, 2637, 2712, 2712, 2746, 2752, 2797, 2808,2860, 2861, 2876,2902,2909, 3010, 3011, 3019, 3079, 3220, 3222, 3237,3285, 3311, 3475, 3581, 3619, 3630, 3662,3679, 3699,3760, 3782, 3785, 3788, 3791,3797, 3799, 3801, 3811, 3815, 3820, 3861, 3868, 3874, 3891, 3904, 3915, 3931, 3936, 3938, 3971, 3980, 3989, 4010, 4012, 4013, 4016, 4020, 4021, 4022, 4027, 4051, 4071, 4102, 4136, 4151, 4191, 4194, 4202, 4208, 4214, 4215, 4216, 4217, 4218, 4220, 4224, 4227, 4231, 4247,4263, 4267,4274, 4277, 4291, 4294, 4295, 4301, 4308, 4320, 4330, 4357, 4374, 4376, 4390, 4391, 4393, 4394, 4407, 4420, 4434, 4464, 4531, 4538, 4603, 4611, 4612, 4640, 4711, 4721, 4751, 4792, 4892, 4910, 4929, 4940, 4990, 5088, 5136, 5210, 5211, 5251, 5279, 5321, 5322, 5324, 5341, 5344, 5350, 5390, 5411, 5414, 5416,5421, 5470,5471, 5472,5499, 5523, 5526, 5529, 5551, 5552, 5579,5587, 5588, 5596, 5751, 5774, 5790, 5801, 5811, 5819, 5871, 5878, 5891, 5906, 5917, 5937, 5963, 5972, 6038, 6071, 6074, 6110, 6112, 6154, 6190, 6196, 6199, 6212, 6220, 6227, 6225, 6257, 6289, 6296, 6310, 6328, 6335, 6384, 6394, 6420, 6447, 6501, 6514, 6520, 6546, 6562, 6565, 6584, 6611, 6615, 6668, 6685, 6699, 6710, 6716, 6726, 6737, 6741, 6797, 6811, 6914, 6916, 6984, 6994, 7015, 7022, 7039, 7135, 7208, 7251, 7344, 7402, 7408, 7414, 7431, 7436, 7474, 7497, 7506, 7511, 7529, 7571, 7594, 7610, 7621, 7631, 7655, 7743, 7746, 7761, 7781, 7791, 7813, 7814, 7820, 7821,7850, 7861, 7867, 7870, 7871, 7880, 7904, 7926, 7934, 7988

Although in the above only 331 tripetides were found to have non zero values the rough set implemented through RSES software gave informative tripeptide number as 345 due to presence of non zero values in other proteins of CCR2. The overall result showed more than 5% increase in accuracy for three (3) of the classes of CC Chemokine GPCRs. The Grid Diagram showing the values of C and Gamma and accuracy with reducts obtained after application of Rough Sets is shown below.

The improvement in results with the application of rough sets in all the cases show the applicability and usage of redundancy removal as a means to improve accuracy deduce the most informative characters. For most of the classes more than 5% increase in accuracy was observed which indicates utility of dimensionality reduction in improving classification accuracy

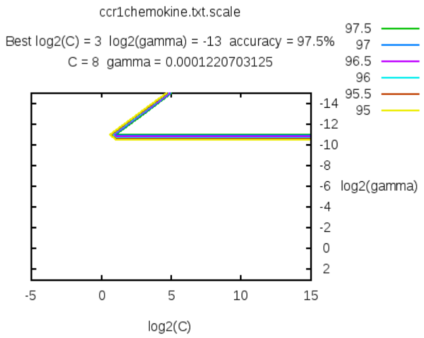

Figure 1. Result with CCR1

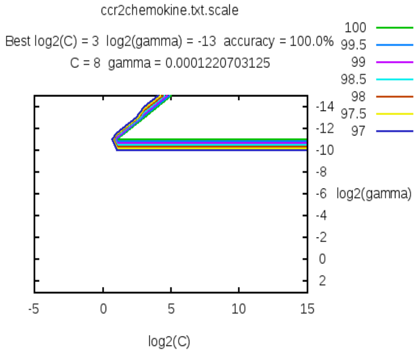

Figure 2. Results with CCR2

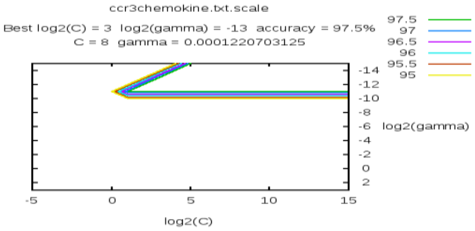

Figure 3. Result with CCR3

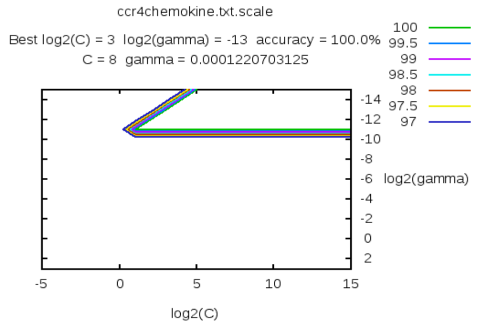

Figure 4. Result with CCR4

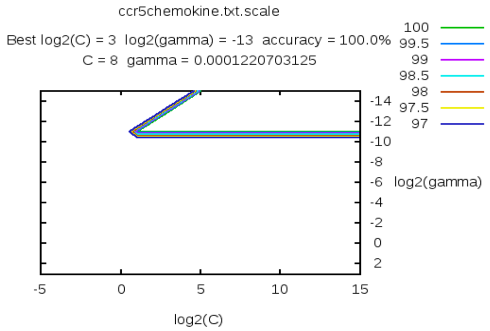

Figure 5. Result with CCR5

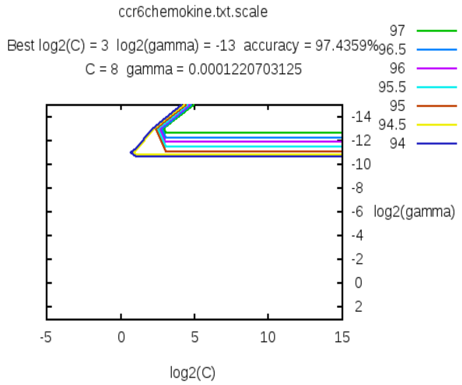

Figure 6. Result with CCR7

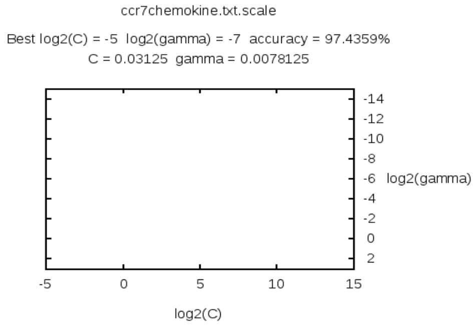

Figure 7. Result with CCR7

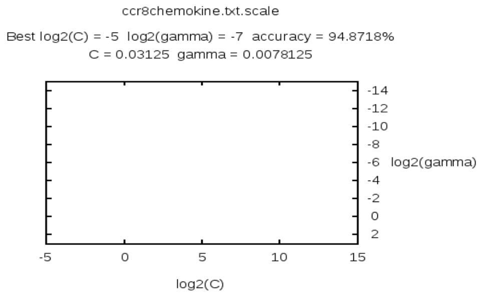

Figure 8. Result with CCR8

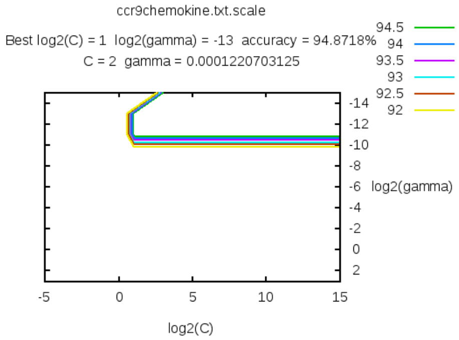

Figure 9. Result with CCR9

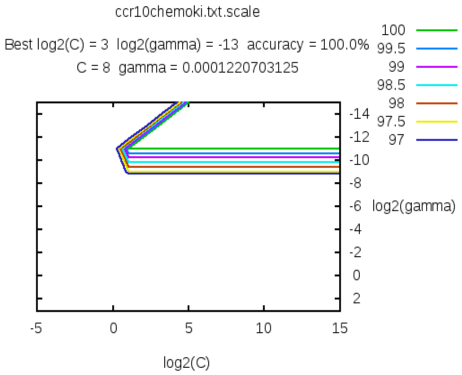

Figure 10. Result with CCR10

CONCLUSION

This is novel concept where by rough set theory has been implemented for dimensionality reduction for tripeptide composition. This concept has wide applicability in data with larger dimensions like tetra peptide composition. More over the reducts obtained through it can be used to build hybrid vectors which can yield even more informative results

REFERENCES

[1] Senselab (http:/senselab.med.yale.edu/NeuronDB/receptors2.asp#Chemokine%20 receptors,CC).

[2] A.P.M.P. Marino, A.A. Silva, “CC-chemokine receptors: a potential therapeutic target for Trypanosoma cruzi-elicited myocarditis,” Mem Inst Oswaldo Cruz, Rio de Janeiro, Vol. 100(Suppl. I): 93-96, 2005.

[3] K.R. Bracke, I.K. Demedts, G.F. Joos, G.G. Brusselle, “CC-chemokine receptors in chronic obstructive pulmonary disease,” Inflamm Allergy Drug Targets. 6(2):75-9, 2005.

[4] S. Segerer, M. Mack, H. Regele, D. Kerjaschki, D. Schlöndorff, “Expression of the C-C chemokine receptor 5 in human kidney diseases,” Kidney Int. 56(1):52-64, 1995.

[5] N. K. Mishra, M. Kumar, , and G.P.S. Raghava, “Support Vector Machine Based Prediction of Glutathine S-Transferase Proteins,” Proteins & Peptide Letters, 14, 575-580, 2007.

[6] RSES 2.2 User’s Guide, Warsaw University http://logic.mimuw.edu.pl/~rses.

[7] D.K. Srivastava, K.S. Patnaik and L. Bhambhu, “Data Classification: A Rough - SVM Approach. Contemporary Engineering Sciences,” Vol. 3, 2010, no. 2, pp77 - 86 (8), 2010.

[8] UniProtKB at www.uniprot.org

[9] Z. Pawlak, “Rough sets: Theoretical aspects of reasoning about data,” Dordrecht: Kluwer, 1991.

[10] Wang Xiaosheng and Gotoh Osamu, “Microarray-Based Cancer Prediction Using Soft Computing Approach,” Cancer Inform., 7: 123–139, 2009.

[11] UniProtKB at www.uniprot.org.

[12] V.Vapnik, “The Nature of Statistical Learning Theory,” Springer Verlag, 1995.

[13] T. Joachims, “Learning to Classify Text Using Support Vector Machines,” Dissertation, Kluwer, 2002.

[14] T. Joachims, “In: Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning,” B. Schölkopf and C. Burges and A. Smola (ed.), MIT Press, 1999.