Granular Support Vector Machine Based Method for Prediction of Solubility of Proteins on Overexpression in Escherichia Coli

Pankaj Kumar[2], V.K. Jayaraman[1,⋆], and B.D. Kulkarni[1,⋆]

[1] Chemical Engineering Division, National Chemical Laboratory, Pune-411008, India
vk.jayaraman@ncl.res.in, bd.kulkarni@ncl.res.in
[2] Department of Chemical Engineering, Indian Institute of Technology, Kharagpur-721302, India

Abstract. We employed a granular support vector Machines(GSVM) for prediction of soluble proteins on over expression in *Escherichia coli* . Granular computing splits the feature space into a set of subspaces (or information granules) such as classes, subsets, clusters and intervals [14]. By the principle of divide and conquer it decomposes a bigger complex problem into smaller and computationally simpler problems. Each of the granules is then solved independently and all the results are aggregated to form the final solution. For the purpose of granulation association rules was employed. The results indicate that a difficult imbalanced classification problem can be successfully solved by employing GSVM.

1 Introduction

The enteric bacterium *Escherichia coli* is the most commonly used organism for the production of recombinant proteins. *E coli* has been preferred over other expression hosts because it is well characterized, is easy to handle and manipulate genetically, and has a relatively high growth and production rate [3]. However only some proteins are soluble upon overexpression in *E coli* and others are generally expressed as insoluble aggregate folding intermediates known as inclusion bodies [3]. It has been observed that primary sequence of the protein is the most important determinant of the solubility status of the overexpressed protein [12]. Protein sequences are difficult to understand and model because of their random length; furthermore solubility of protein on overexpression in *E coli* is manifestation of the net effect of several sequence dependent and sequence independent factors [11]. Wilkinson and Harrison [18] observed that inclusion body formation is correlated, in descending order, to charge average, turn forming residue fraction, cysteine fraction, proline fraction, hydrophobicity and molecular weight. But later it was found by Davis *et al*, 1999 that only the first two features are critical in distinguishing soluble and insoluble proteins. This problem was further investigated by several authors [4,7,15] etc.

[⋆] Corresponding author.

A. Ghosh, R.K. De, and S.K. Pal (Eds.): PReMI 2007, LNCS 4815, pp. 406–415, 2007.

Aliphatic index, the frequency of occurrence of Asn, Thr and Tyr and the dipeptide and tripeptide-composition were found to be the most informative features by Idiculla Thomas and Balaji, 2005. Recently, Idicula-Thomas *et al*, 2006 employed Support Vector Machines(SVM) to predict solubility on overexpression. As the data was unbalanced they had employed the weighted version of the SVM which yielded an accuracy of 72% with a specificity of 76% and sensitivity of 55 %. The algorithm could satisfactorily predict the change in solubility for most of the point mutations reported in literature. Due to the immense importance of this classification problem, it would be highly desirable to increase the prediction accuracy and the sensitivity. In the present work a granular computing based machine learning approach has been employed with a view to improve the prediction performance. Granular computing, unlike traditional, computing, is knowledge oriented. In this work association rules have been used to make granules while SVM is employed as a classifying method. Our result shows superiority of the proposed method over building a single contiguous hyperplane to classify the data using SVM.

2 System and Methods

In this section a brief introduction of the principle of granular computing and support vector machine is presented. Subsequently the methodology employed in building Granular Support vector machines (GSVM) is explained. GSVM combines statistical machine learning algorithm with knowledge-based classification to build a robust model.

2.1 Granular Computing

Granular computing('GrC') was first introduced by T.Y.Lin, in 1997. Since then it has been successfully employed in various fields like diakoptics, divide and conquer structured programming, interval computing, cluster analysis, fuzzy and rough set theories, neutrosophic computing, quotient space theory, belief functions, machine learning, databases, and many others. [9,14,19].

Granular computing splits the feature space into a set of subspaces (or information granules) such as classes, subsets, clusters and intervals [14]. By the principle of divide and conquer it decomposes a bigger complex problem into smaller and computationally simpler problems. Each of the granules is then solved independently and all the results are aggregated to form the final solution. Proper granulation is capable of removing some redundant and irrelevant information and at the same time facilitates getting rid of overfitting problem [16]. Thus granulation helps in building a computationally more efficient model for a complex problem.

In granular computing information granules are first constructed and computations are subsequently carried out with the granules [19] In the literature several methods have been used for granulation like clustering [21], fuzzy sets [20], decision tree and association rules. In this work association rules [16] have been used for the purpose of granulation.

2.2 Association Rules

Association rules tend to capture the underlying hidden patterns in datasets [1]. It provides information in the form of "IF - THEN" statements. In the most general form an association rule has the form IF C_1 THEN C_2 where C_1 and C_2 are conjunctions of condition and each condition is of form either $A_i = V_i$ or $A_i \in (L_i, U_i)$. For e.g., IF *frequency of occurrence of Cysteine* (Cys) lies between 0.024 and 0.3279 THEN *that protein belongs to class -1*(inclusion bodies) The antecedent part ('IF part') can have one or more than one condition joined by an operator *and*. It must, however, be beneficial to form association rules with short IF-parts to avoid overfitting, yielding better generalization. To estimate the quality of a rule formed the confidence and support parameters are used:

Confidence. Confidence is defined as the fraction of instances that are correctly classified by the rule among the instances for which it makes any prediction. Thus if the confidence of a rule is one, we can say that all the data in the training sample that satisfies the rule are correctly classified. Mathematically confidence for an association rule can be represented as,

$$con = \frac{count\_then}{count\_if}$$

Where, *count_then*, is the number of sample that satisfies the $THEN$ part of the rule, and *count_if*, is the number of samples that satisfy only the IF part of association rule.

Support. Support is defined as the ratio of the training data that are correctly classified by the rule to the size of training data with same class label as then part. Hence a support indicates the fraction of data in a class correctly classified by the rule:

$$\sup = \frac{count\_then}{size(class_{then})}$$

Where, $size(class_{then})$ is the size of training data with same class label as rule consequent.

While making a rule a threshold for support and confidence is used to prune out all the rules that will have support and confidence below the user defined threshold [2]. This is done so as to obtain a set of association rules that will enable efficient and reliable classification of unseen test instances. If the threshold confidence of the rule is kept high then less number of association rules will be mined but their prediction accuracy will be quite high. Similarly if support is kept very high generalization will be more but number of rules obtained will be very few while if support is too low then rule obtained will tend to overfit the training sample.

2.3 Support Vector Machines(SVM)

SVM are a machine learning algorithm introduced by Vapnik [17].It classifies a nonlinearly separable problem by building a linear separating hyperplane in a high dimensional feature space. The general hyperplane equation is of the form $w^T.x + b = 0$ where w is the weight vector, b is the bias and (.) denotes the dot product. Here w and b are selected so as to maximize the margin,$\frac{1}{\|w\|}$ between the hyperplane and the closest data points belonging to the different classes. The computational intractability problem introduced due to the high dimensionality is over come by defining an appropriate Kernel function.

Given a training dataset of the form (x_i, y_i) , $i = 1, 2, ..., N$ where $x_i \in \mathbb{R}^m$ and $y_i \in \{-1, 1\}$, the final SVM classifier function can be given as:

$$f(x) = \sum_{i=1}^{m} y_i \alpha_i K(x_i, x) + b \qquad (1)$$

Here x_i represent i^{th} vector of input pattern and y_i is the target output corresponding to the i^{th} vector and , K is the kernel matrix and $m(N)$ is the number of input pattern having non zero Langrangian multipliers(α_i); these are also called support vectors. The Langrangian multiplier is found by solving the following dual form of quadratic programming problem,

$$w(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \qquad (2)$$

Subject to the constraint

$$0 \leq \alpha_i \leq C, i = 1, 2......N$$

and

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

Where C is the regularization parameter known as cost function that determines the tradeoff between the model complexity and the misclassification. For imbalanced classification problem SVM uses different error cost for the positive C^+ and negative C^- classes. Here the langrangian equation is modified to

$$L_p = \frac{\|w\|^2}{2} + C^+ \sum_{i|y_i=+!}^{n_+} \xi_i + C^- \sum_{j|y_j=-!}^{n_-} \xi_j - \sum_{i=1}^{n} \alpha_i \left[y_i \left(w.x_i + b \right) - 1 + \xi_i \right] - \sum_{i=1}^{n} r_i.\xi_i$$

Subject to the constraint

$$0 \leq \alpha_i \leq C^+ \mathrm{if y_i = +1, and if y_i = -1}$$

After the optimal value of α^i is found the decision function is based on the sign of $f(x)$ as given by eq.(1).

Different types of kernel function (Burges 1998) are used for transformation of input space to a higher dimension feature space .Most commonly used kernel function are

Linear: $K(x_i, x) = x_i^T x_j$

Polynomial: $K(x_i, x) = (\gamma x_i^T x_j + r)^d$

Radial bias function: $K(x_i, x) = \exp(-\gamma \left\| x_i - x_j \right\|^2)$

In the present work RBF kernel was found to provide the best possible results.

2.4 Granular Support Vector Machine (GSVM)

For many complex problems it can never be guaranteed that a contiguous hyperplane as discussed above will be able to classify the data correctly. Better classification performance can be achieved by judicious granulation of the feature space. For example, consider the traditional XOR problem; as such without any transformation it is non-linearly separable but if we divide the whole feature into two equal halves then each half becomes linearly separable. Even in the case where a single linear hyperplane is available the use of granulation will help in maximizing the margin between the hyperplane and the closest data points belonging to the different classes.

Furthermore for the case of imbalanced data where the number of instances in one class is far more than the number of instances in other, the separating hyperplane tends to shifts towards the minority class so SVM misclassifies most of the instances into the majority class, thus giving higher accuracy for the majority class but poor predictivity for minority class. In such cases granulation may become an effective means to handle data imbalance. In GSVM the whole feature space is first divided into granules, *viz.*, pure (where almost all the instances belong to one class) and mixed granules (where instances from both classes are present). After separating out pure granules instances present in the mixed granule may become more balanced and hence the probability of prediction accuracy of SVM can be expected to be higher.

3 Modeling Method

3.1 Association Rules Formation as Granulation Methodology

In this work we employed the association rules methodology of Tang *et al., 2005* for the purpose of granulation. As explained in section (2.2) association rules with optimal support and confidence were mined. Among all the association rules formed with different attributes the one with highest confidence was added to the set called *selected_rule_set*. After a rule was chosen, all instances classified by the rule were removed and the rule formation process was repeated until no further rule with support and confidence greater than the predetermined threshold is formed or all the instances has already been classified. Care was taken to apply them in the order in which they are discovered. Table 1 shows the *selected_rule_set* for dataset with all 446 features.

3.2 GSVM Modeling

After all the association rules were obtained GSVM model was built by iteratively combining the association rules from *selected_rule_set* to find the optimal granules which were both pure and significant. Thus using association rules the complete feature space can be divided into three different granules[16]:

Positive pure granule (PPG) in which almost all the data belong to positive class.

Negative pure granule (NPG) in which almost all the instances belong to negative class.

Mixed zone (MG) or mixed granule which contains instances belonging to both the classes

To begin with, over the complete feature space, cross validation performance of SVM in the training dataset was obtained and was taken as baseline accuracy. The subsequent algorithmic steps in the GSVM model are:

1. A rule from *selected_rule_set* was taken

2. All the instances that satisfy the antecedent (*If part*) part of rule was removed as pure granules and was assigned the class label as predicted by consequent part of rule.

3. the remaining instances that do not satisfy the rule, form the mixed granule. SVM model was built with the instances in the mixed granule by tuning the algorithm parameters to obtain the best accuracy.

4. If the considered rule was added to set called *final_rule_set* if it was found to improve the classification performance. The improved accuracy was now considered to be the new baseline accuracy.

5. Otherwise the next rule in the list from *selected_rule_set* was taken and steps 3 to 5 were repeated.

6. The above steps were continued until the entire set of rules in the list had been processed.

Table 2 shows the *final_rule_set* for the dataset comprising of all 446 features. When unseen test instances are to be classified, they are first checked by the formed association rules in *final_rule_set*. All the instances that satisfy the antecedent of rule are assigned the class predicted by the rule and the class label for the remaining instances (not predicted by rule) were predicted by the SVM model built on mixed granule.

4 Experimental Evaluation

4.1 Data Description

The Dataset of Idicula-Thomas et al. [12] were employed for the GSVM experiments. This dataset consist of 192 protein sequences, 62 of which are soluble on overexpression in *E.Coli* and the remaining 130 sequences form inclusion body. The 446 features extracted by them include i) six physiochemical properties(Attribute nos. 1-6), *viz.*, aliphatic index, instability index of the entire protein, instability index of the N-terminus and net charge. ii) twenty single aminoacid residues(Attribute nos. 7-26) arranged in alphabetical order (A,C,D) followed by 20 reduced alphabets(attribur nos. 27-46). The reduced alphabets employed includes 7 reduced class of conformational similarity,8 reduced class of BLOSUM50 substitution matrix and 5 reduced class of hydrophobicity [12]. Finally the features in the list includes 400 attributes(attribute nos. 47-446) comprising of the dipeptide compositions.

4.2 Model Building

The instances(each comprising of 446 features)were randomly divided into training and test sets keeping the inclusion body forming and the soluble proteins approximately in ratio of 2:1. The training dataset comprised 128 sequences, 87 inclusion body-forming and 41 soluble proteins. The test dataset comprised 64 sequences, 43 inclusion body forming and 21 soluble proteins. [12] The modeling process was initiated by first forming association rules with the instances in the training dataset. As explained in section (3.1), only single feature association rules with substantial support and confidence were mined to form *selected_rule_set*. Table 1 shows the mined set of association rules in the form: . IF $X_0 \leq attribute_i \leq X_1$ THEN $class = y$

Table 1. Mined association rule on original unscaled training data

S.No.	X_0	X_1	Attribute Number	Confidence	Support	Class
1	0.0051	0.0242	443	1	0.1954	-1
2	0.0059	Inf	435	1	0.1609	-1
3	0.0078	0.0127	296	1	0.1609	-1
4	0.0084	0.0141	330	0.9231	0.2927	1

GSVM model was built employing these rules for pure zones and SVM classification for the mixed zone. However before applying SVM, as a preprocessing step all the features were scaled by making their mean zero and standard deviation one. SVM experiments done in this work were performed using an implementation of LIBSVM (chang Lin, 2001). As our data was imbalanced weighted SVM was used. The SVM parameters C,γ and weights were tuned by grid search. Table 2 shows the final set of rule selected by GSVM algorithm to make a model. Out of the 4 rules shown in table 1 only the rules shown in table 2 were found to increase the cross-validation performance over training data, so only those two rules were selected.

The algorithm performance was subsequently tested on unseen test dataset using the same test measure as used by Idicula-Thomas and Kulkarni et al. [12]. 50 random splits of the dataset were taken (with the same ratio of nearly 1:2 between the two classes of proteins), and their average performance was measured. Table 3 shows the comparison of results obtained by using GSVM and SVM (as reported by [12]). These results shows that the GSVM is capable

Table 2. Final set of rules selected by GSVM algorithm

S.No.	X_0	X_1	Attribute Number	Confidence	Support	Class
1	0.0078	0.127	296	1	0.1609	-1
2	0.0084	0.141	330	0.9231	0.2927	1

Table 3. Classification result on test dataset averaged over 50 random splits

Number of features	Algorithm	ROC	Accuracy (%)	Specificity(%)	Sensitivity(%)
446	SVM	0.5316	72	76	55
446	GSVM	0.7227	75.41	81.40	63.14
27	GSVM	0.7635	79.22	84.70	68

of capturing inherent data distribution more accurately as compared to a single SVM build over complete feature space.

As the number of proteins forming inclusion bodies is far more than number of soluble proteins(nearly 2 times,) our dataset is imbalanced. So accuracy alone does not give the correct measure of performance. For an imbalanced data, receiver operation characteristic (ROC) curve is generally used as test measure. Our result shows a marked increase in the value of ROC from 0.5316 using SVM over complete feature to 0.72227 using GSVM for the *bestclassifier* reported by Idicula-Thomas A.J.Kulkarni *et al.*, 2006 using 446 features. The value of sensitivity and specificity has also gone up which has increased the overall accuracy to 75.41%. The increased ROC shows that our model is not biased towards majority class and is capable of predicting the minority class (soluble proteins) as well with equally good accuracy.

We also tried feature selection in the mixed granule with the original 446 features to find the most informative subset. After feature selection only 27 features were found critical for predicting the solubility. The selected features were aliphatic index, frequency of occurrence of residues Cysteine (Cys), Glutanic acid (Glu), Asparagine (Asn) and Tyrosine (Tyr). Among the reduced alphabets, only the reduced class [CMQLEKRA] was selected from the seven reduced classes of conformational similarity. Similarly from the five reduced classes of hydrophobicity originally reported, only [CFILMVW] and [NQSTY] were selected. And from the eight reduced classes of BLOSUM50 substitution matrix the only reduced class selected was [CILMV]. The 18 dipeptide whose composition were found to significant. These include [VC], [AE], [VE], [WF], [YF], [AG], [FG], [WG], [HH], [MI], [HK], [KN], [KP], [ER], [YS], [RV], [KY], and [TY]. A new GSVM model was built with these most informative features.

In this case we didn't get any positive rule, which satisfied our minimum support and confidence threshold condition (kept as 0.18 and 0.85 respectively for positive rule while for negative rule the values are 0.15 and 0.95 respectively). After applying GSVM all 3 rules were selected in the final model. So our final

Table 4. Mined association rule on original unscaled training data after feature selection

S.No.	X_0	X_1	Attribute Number	Confidence	Support	Class
1	0.0059	Inf	26	1	0.1954	-1
2	0.0240	0.3279	2	1	0.1609	-1
3	0.0027	0.3279	12	1	0.1954	-1

model with 27 selected features comprised of association rules shown in table 4 and SVM parameters C=32, $\gamma = 0.0039$ and W=1.3. Our result (Table 3) shows that performance was further improved by feature selection.

5 Conclusion

In this work Granular Support Vector Machines(GSVM) was successfully employed for classification of soluble and insoluble proteins. GSVM systematically combines statistical learning theory with granular computing to build a hybrid system exhibiting superior performance with the inherently unbalanced data set. By splitting the feature space into granules it reduces the complexity of problem thereby improving the classification efficiency. The significant increase in ROC values as compared to that obtained using SVM alone bears testimony to the excellent generalization capability of the hybrid model.

Acknowledgement

The authors are grateful to Department of Biotechnology (DBT) for providing financial assistance.

References

1. Agrawal, et al.: Mining association rules between sets of items in large databases. In: Proceedings of the ACM SIGMOD Conference on Management of Data, Washington, D.C., pp. 207–216 (May 1993)
2. Agrawal, R., Ramakrishnan, S.: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, pp. 12–15. Morgan Kaufmann, San Francisco (1994)
3. Baneyx, F.: Recombinant protein expression in Escherichia coli. Curr. Opin. Biotechnol. 10, 411–421 (1999)
4. Bertone, P., et al.: SPINE: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics. Nucleic Acids Res. 29, 2884–2898 (2001)
5. Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining Knowledge Disc 2(2), 121–167 (1998)
6. Davis, G.D., Elisee, C., Newham, D.M., Harrison, R.G.: New Fusion Protein Systems Designed to Give Soluble Expression in Escherichia coli. Biotechnol. Bioeng. 65, 382–388 (1999)

7. Goh, C.S., et al.: Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. J. Mol. Biol. 336, 115–130 (2004)
8. Harrison, R.G.: Expression of soluble heterologous proteins via fusion with NusA protein. inNovations 11, 4–7 (2000)
9. Hirota, K., Pedrycz, W.: Fuzzy computing for data mining. Proceedings of the IEEE 87, 1575–1600 (1999)
10. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
11. Idicula-Thomas, S., Balaji, P.V.: Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. emphProtein Sci. 14, 582–592 (2005)
12. Idicula-Thomas, S., Kulkarni, A.J., Kulkarni, B.D., Jayaraman, V.K., Balaji, P.V.: A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. Bioinformatics 22, 278–284 (2006)
13. Keerthi, S.S., Lin, C.-J.: Asymptotic behaviors of support vector machines with Gaussian kernel. Neural Computation 15(7), 1667–1689 (2003)
14. Lin, T.Y.: Granular computing, Announcement of the BISC Special Interest Group on Granular Computing (1997)
15. Luan, C.H., et al.: High-throughput expression of C. elegans proteins. Genome Res. 14, 2102–2110 (2004)
16. Yuchun, T., Bo, J., Zhang, Y.-Q.: Granular support vector machines with association rules mining for protein homology prediction, Artificial Intelligence in Medicine. Computational Intelligence Techniques in Bioinformatics 35(1-2), 121–134 (2005)
17. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, New York (1995)
18. Wilkinson, D.L., Harrison, R.G.: Predicting the solubility of recombinant proteins in Escherichia coli. Biotechnology 9, 443–448 (1991)
19. Yao, Y.Y.: Granular computing: basic issues and possible solutions. In: Wang, P.P. (ed.) Proceedings of the 5th Joint Conference on Information Sciences, Atlantic City, New Jersey, USA. Association for Intelligent Machinery, vol. I, pp. 186–189 (2000)
20. Zadeh, L.A.: Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems 90(2), 111–127 (1997)
21. Zhong, W., He, J., Harrison, R., Tai, P.C., Pan, Y.: Clustering support vector machines for protein local structure prediction. Expert Systems with Applications 32(2), 518–526 (2007)