ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Misclassification of class C G-protein-coupled receptors as a label noise problem

Caroline König[1], Alfredo Vellido[1], René Alquezar[1,2] and Jesús Giraldo[3] *

1- Dept. de Llenguatges i Sistemes Informàtics, Univ. Politècnica de Catalunya
C. Jordi Girona, 1-3, 08034, Barcelona - Spain

2- Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona - Spain

3- Institut de Neurociències - Unitat de Bioestadística
Univ. Autònoma de Barcelona, 08193, Cerdanyola del Vallès (Barcelona) - Spain

Abstract. G-Protein-Coupled Receptors (GPCRs) are cell membrane proteins of relevance to biology and pharmacology. Their supervised classification in subtypes is hampered by label noise, which stems from a combination of expert knowledge limitations and lack of clear correspondence between labels and different representations of the protein primary sequences. In this brief study, we describe a systematic approach to the analysis of GPCR misclassifications using Support Vector Machines and use it to assist the discovery of database labeling quality problems and investigate the extent to which GPCR sequence physicochemical transformations reflect GPCR subtype labeling. The proposed approach could enable a filtering approach to the label noise problem.

1 Introduction

Machine learning (ML) is a data-driven process and, as such, the quality of the available data is paramount. Label noise may become a data quality problem in supervised ML and is commonplace in real-world applications. It can take many forms, including expert subjectivity in the labelling process, bounds on the available information and communication noise [1].

There are few domains of knowledge in which the effects of label noise are so pervasive and eloquent as in biomedicine and bioinformatics [2]. In medicine, for instance, the reliability of diagnostic labels is often bounded by the natural limitations of the specialists' expertise [3], or even by the formal requirement of a consensual or majority-based decision-making procedure.

In bioinformatics, protein subtype characterization is riddled with this problem. In the specific area of G-Protein-Coupled Receptors (GPCRs), this problem is magnified by the fact that subtyping can be performed at up to seven levels of detail [4]. GPCRs are cell membrane proteins of relevance to biology due to their role in transducing extracellular signals and to the pharmaceutical industry for being the target for many new therapies in pain, anxiety and neurodegenerative disorders, amongst others.

Our study focuses on the characterization of class C, one of the five GPCR families. The 3-D structure of proteins is key to the determination of their

*This research was partially funded by Spanish MINECO TIN2012-31377 project

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

function but, so far, no class C full 3-D structure has yet been discovered and their functional study must mostly rely on primary structure: the amino acid (AA) sequences, publicly available from several databases. There are seven class C subtypes with their corresponding labels. Label noise is unavoidable in this context because sequence labeling is itself model-based and follows a complex many-step procedure that can only guarantee limited success [5]. GPCR classification may use aligned or unaligned versions of sequences. Some methods of sequence alignment-free analysis entail transforming sequences according to the physicochemical properties of their constituent AAs [6].

In this study, we investigated the classification of several alignment-free transformations of class C GPCR sequences using Support Vector Machines (SVM). The misclassified sequences were analyzed to discover non-random label noise effects as a way to explore their possible biological explanation. This could be the proof of concept for a systematic approach to assist the discovery of GPCR database labelling quality problems, which would become the core of a label filtering decision support system [1].

2 Materials

The data analyzed in this study were extracted from GPCRDB[1][5], a curated and publicly accessible database of GPCRs. The investigated dataset (version 11.3.4 as of March 2011) comprises a total of 1,510 class C GPCR sequences, belonging to seven subfamilies and including: 351 metabotropic glutamate (mG), 48 calcium sensing (CS), 208 GABA-B (GB), 344 vomeronasal (VN), 392 pheromone (Ph), 102 odorant (Od) and 65 taste (Ta).

3 Experiments

Previous research [7] investigated the supervised classification of the data set described in section 2 using different classifiers for different alignment-free transformations of the sequences, including AA composition (AAC), digram-frequency composition (Digram), Auto-Cross Covariance (ACC) [8] and the Physicochemical Distance-Based Transformation (PDBT) [6]. AAC and Digram measure, in turn, the frequency of appearance of N-grams of length one and two in the sequence, while ACC and PDBT are more complex transformations based on the physicochemical properties of the AAs and the sequencing information. Table 1 shows the best classification results, obtained with SVM, for the different transformed data sets.

The detailed analysis of the per-class results revealed relatively minor differences between those obtained with each of the four transformed data sets. This observation suggested that the main causes of misclassification lie beyond the differences between data transformations and that a more systematic analysis of the classification errors was required.

[1] `http://www.gpcr.org/7tm`

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

Data	**Accu**	**MCC**
AAC	0.88	0.84
Digram	0.93	0.91
ACC	0.93	0.91
PDBT	0.92	0.90

Class	**MCC**	**Prec**	**Rec**
mG	0.95	0.95	0.99
CS	0.93	1.00	0.88
GB	0.98	0.99	0.99
VN	0.89	0.91	0.92
Ph	0.86	0.89	0.90
Od	0.79	0.89	0.74
Ta	0.99	1.00	0.98

Table 1: SVM classifier results; Left: Global results for the four data transformations; accuracy (Accu), Matthews Correlation Coefficient (MCC). Right: Class C GPCR subtype-specific results for the ACC data set only, including MCC, Precision (Prec) and Recall (Rec).

♯	ER	TC	**mG**	**CS**	**GB**	**VN**	**Ph**	**Od**	**Ta**	VT	VP	R
2	100	CS	100	0	0	0	0	0	0	91	600	0.15
6	100	VN	0	0	0	0	96	4	0	404	596	0.67
7	100	VN	100	0	0	0	0	0	0	300	600	0.5

Table 2: Example of misclassification statistics for the ACC data set. For each sequence ♯, the error rate (ER), the true class (TC), and how many times this sequence was misclassified as belonging to each of the other classes (mG-Ta), are displayed. The three last columns display the sum of the votes for the true class (VT), for the most frequently predicted class (VP), and the ratio (R) of one to the other.

3.1 A systematic approach to GPCR misclassification analysis

3.1.1 Iterative classification with different classification models

The proposed approach entails repeating 100 times the following procedure: First, using 5-cross validation (5-CV), so that the current training set is used to construct a RBF-SVM model [9] with an optimal value for the γ parameter of the kernel function and with the error penalty parameter C varying within a small range near its previously established optimum value; then classifying the test set, recording which GPCR sequences are misclassified and the corresponding confusion matrix. The use of CV in each of the 100 iterations ensures that each instance is used one time for classification in each iteration of the outer loop.

We now have detailed results of how many times a sequence was misclassified and how many times it was assigned to another class. To focus on the most consistent classification errors, we set a conservative misclassification boundary of 75% (i.e., only sequences misclassified in at least a 75% of occasions are deemed to be misclassifications). Table 2 shows some examples for the ACC data set. See Table 3 for the mapping between the number ♯ and the protein database *Id.*

This misclassification analysis was repeated for each of the transformed data sets. The AAC, Digram, ACC and PDBT sets yielded, in turn, 143, 88, 85

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

and 100 misclassifications. A detailed analysis of these frequently misclassified sequences revealed that they are nearly identical for ACC and Digram. There are some differences with the PDBT misclassifications that might be the result of the very different type of transformation. Importantly, there are 53 frequently misclassified sequences that are common to all four data sets.

3.1.2 Analysis of misclassifications according to the voting scheme

These results suggest the existence of subtypes with recurrently wrong class assignments. Since the underlying classification scheme of the SVM implementation [9] was "one-vs-one", we first decided to analyze the results of the voting scheme as applied to the $k(k-1)/2$ resulting classifiers, including the votes of each one, for each instance in each iteration. According to libSVM[2], the subtype with the greatest number of votes becomes the predicted class.

When the *voting ratio* (R) of true class to predicted class is low (≤ 0.5), the classification error was deemed to be *large*, and *small* otherwise. To illustrate this, we show the voting scheme results for the selected instances of Table 2. Sequence ♯ 6, for instance, is a VN consistently misclassified as Ph; the magnitude of the error is small, though, as R=0.67>0.5 is high. Sequence ♯ 2 is a CS, consistently misclassified as mG; the magnitude of the error is large, as R=0.15≤0.5 is low.

Only 7 of the 85 frequently misclassified ACC-transformed sequences yield large errors. Similarly, for AAC, Digram and PDBT sets, the majority of sequences have small errors.

3.1.3 Analysis of misclassifications according to the decision values

Clear differences in the magnitude of the recurrent classification errors have been found. Pursuing further insight, we define a *cumulative decision value* (CDV) specifically for the binary classifier that involves the true class and the predicted class. The CDV is calculated as the sum of the SVM decision values over all iterations and was recorded for each instance. GPCR subtypes were numbered 1 to 7 in the order they are presented in section 2. For subtypes i, j, if i is the true class, a large positive CDV value if $i > j$ and a large negative one if $i < j$ both indicate clear misclassifications.

This time, the magnitude of the error was deemed *large* or *small* depending on whether the CDV exceeded the threshold of 60 in absolute value or not. A total of 21 out of the 85 frequently misclassified instances of the ACC transformed data set have a *large* error according to this criterion, whereof 4 yield a *very large* one (≥ 95).

Note that the information conveyed by the CDV complements that of R. For instance, a misclassified sequence with high R would suggest that voting discards all subtypes but the true and the predicted ones. If this is accompanied by a large CDV in absolute value, the predicted subtype is strongly preferred.

[2]http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

♯	Id	TC	PC	R	CDV
1	q5i5c3_9tele	mG	Od	0.75	-95
2	XP_002123664	CS	mG	0.15	50
3	q8c0m6_mouse	CS	Ph	0.15	-46
4	XP_002740613	CS	mG	0	-66
5	XP_002936197	VN	Ph	0.83	-96
6	XP_002940476	VN	Ph	0.67	-95
7	XP_002941777	VN	mG	0.5	45
8	B0UYJ3_DANRE	Ph	mG	0.79	109
9	XP_001518611	Od	mG	0.31	46
10	XP_002940324	Od	VN	0.49	70
11	GPC6A_DANRE	Od	Ph	0.5	74

Table 3: Sequences with large classification errors: For each sequence, the GPCRDB Identifier (Id), the true class (TC), the predicted class (PC), the *voting ratio* (R) and the *cumulative decision value* (CDV) are displayed.

4 Discussion

The proposed approach has revealed the existence of a number of instances that, independently of the sequence transformation method, induce certain classification errors that could be deemed large or small (according to criteria that, ultimately, should be set by proteomics experts).

Importantly, this analysis has shown that the misclassifications of a sizeable proportion of sequences have a small magnitude. All these sequences might well be considered as mild cases of label noise and should be redirected to a human expert for further analysis. Small errors also suggest underlying similarities between the GPCR subtypes whose characteristics may be unknown and worth investigating. A small number of instances, though, show consistent and large classification errors. They merit detailed study because they might be affected by a more radical type of label noise, or even by straight mislabelling. In Table 3, we list GPCRs with either very large CDV (4), or small R (7).

Sequences $XP\_002123664$, $XP\_002740613$, $XP\_002936197$, $XP\_002940476$ and $XP\_002940324$ are all recurrently misclassified. $XP\_002740613$, in particular, yields a 100% error ($R = 0$) and large CDV. Their labels should require further expert assessment, given that they were derived by an automated computational analysis from an annotated genomic sequence by means of a gene prediction mode from the RefSeq[3] databank.

Another couple of interesting cases are $q8c0m6\_mouse$ and $B0UYJ3\_DANRE$. According to the information referenced at UniProt[4], these GPCRs are unreviewed and should be considered only as preliminary data. The former, according to GPCRDB, is a CS that our system confidently ($R = 0.15$) classifies as Ph. The European Nucleotide Archive[5] lists it as similar to the putative Ph receptor V2R2. The latter, according to GPCRDB, is a Ph, while our system predicts

[3] http://www.ncbi.nlm.nih.gov/refseq/
[4] http://www.uniprot.org/uniprot/{B0UYJ3,Q8C0M6}
[5] http://www.ebi.ac.uk/ena/data/view/BAC26854

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

it to be an mG with a very large CDV (109). Agreeing with our prediction, the Ensembl Genome Browser[6] considers it to be an mG of subtype $6a$.

5 Conclusions

Label noise is a potentially big problem in the process of automated class C GPCR subtype classification from the alignment-free transformed versions of protein primary sequences. This is because the labels of these sequences are obtained indirectly through complex, many-step similarity modelling processes. In this brief paper, we have proposed a systematic procedure, based on SVM classification, to single out and characterize GPCR sequences with consistent misclassification behaviour. The reported preliminary experimental results are a proof of concept for the viability of a decision support system that combined this procedure with expert knowledge in the field to assist the discovery of GPCR database labelling quality problems, as a basis for label filtering.

References

[1] B. Frénay, G. de Lannoy and M. Verleysen, Label noise-tolerant hidden Markov models for segmentation: application to ECGs. In D. Gunopulos et al. (eds.), Machine Learning and Knowledge Discovery in Databases, LNCS 6911, pages 455–470, Springer, 2011.

[2] P.J.G. Lisboa, A. Vellido and J.D Martín, Computational Intelligence in biomedicine: Some contributions. In M. Verleysen, ed., *procs. of the* 18^{th} *European Symposium on Artificial Neural Networks* (ESANN 2010), d-side, pp. 429-438, Bruges(Belgium), 2010.

[3] A. Vellido, E. Romero, F.F. González-Navarro, L. Belanche-Muñoz, M. Julià-Sapé and C. Arús, Outlier exploration and diagnostic classification of a multi-centre ^{1}H-MRS brain tumour database, *Neurocomputing*, 72(13-15):3085-3097, Elsevier, 2009.

[4] Q.-B. Gao, X.-F. Ye and J. He, Classifying G-Protein-Coupled Receptors to the finest subtype level, *Biochemical and Biophysical Research Communications*, 439(2):303–308, Elsevier, 2013.

[5] B. Vroling, M. Sanders, C. Baakman, A. Borrmann, S. Verhoeven, J. Klomp, L. Oliveira, J. de Vlieg and G. Vriend, GPCRDB: information system for G protein-coupled receptors, *Nucleic Acids Research*, 39(suppl 1):D309-D319, 2011.

[6] B. Liu, X. Wang, Q. Chen, Q. Dong and X. Lan, Using amino acid physicochemical distance transformation for fast protein remote homology detection, *PLoS ONE*, 7(9):e46633, 2012.

[7] C. König, R. Cruz-Barbosa, R. Alquézar and A. Vellido, SVM-based classification of class C GPCRs from alignment-free physicochemical transformations of their sequences. In A. Petrosino, L. Maddalena, P. Pala (Eds.): ICIAP 2013 Workshops, Lecture Notes in Computer Science 8158, pages 336–343, Springer, 2013.

[8] S. Wold, J. Jonsson, M. Sjöström, M. Sandberg and S. Rännar, DNA and peptide sequences and chemical processes multivariately modelled by Principal Component Analysis and Partial Least-Squares projections to latent structures, *Analytica Chimica Acta*, 277:239-253, 1993.

[9] C. Chang and C. Lin. LIBSVM: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, 2011.

[6] `http://www.ensembl.org`