# Corpus-Based vs. Model-Based Selection of Relevant Features

**C. Goutte[†], P.B. Dobrokhotov[*], E. Gaussier[†], A.-L. Veuthey[*]**

[†] *Xerox Research Centre Europe*
*6 ch. de Maupertuis - F-38240 Meylan, France*
*Eric.Gaussier,Cyril.Goutte@xrce.xerox.com*
[*] *Swiss Institute for BioInformatics*
*CMU, 1 Michel-Servet, CH-1211 Genève 4, Switzerland*
*pavel.dobrokhotov@isb-sib.ch,anne-lise.veuthey@isb-sib.ch*

RÉSUMÉ. *Le travail que nous présentons ici a pour but la comparaison de méthodes de sélection d'attributs. Plus précisément, nous nous intéressons à deux grandes approches, celles fondées uniquement sur les données, approche classique qui permet de ne se reposer, pour la construction de modèles de catégorisation, que sur un ensemble restreint, mais pertinent, d'attributs, et celles qui découlent d'un modèle appris. Ces dernières permettent d'expliquer les décisions prises par un modèle, et fournissent aux utilisateurs des moyens de voir ce qui se passe à l'intérieur de la "boîte noire" qu'est bien souvent un catégoriseur. De plus, la comparaison de ces deux approches permet d'évaluer dans quelle mesure les premières sont suffisamment sélectives comparées aux deuxièmes. Notre comparaison expérimentale est réalisée pour une large part sur une collection de résumés médicaux constituée par l'Institut Suisse de Bioinformatique.*

ABSTRACT. *In this contribution, we review a number of approaches to feature selection, divided in two broad classes. Some are corpus-based, ie they use only the data to assess the relevance of each feature, and aim at identifying a small subset of relevant features on which to train categorisation models. Others are model-based, ie they assess the relevance of each feature on the basis of the model used for categorisation. This second class of measures allows to better understand the model decisions. Furthermore, comparing the two classes provide insight on whether or not corpus-based feature extraction is selective enough, and does not overgenerate compared to model-based selection. Our experimental comparison is mainly based on a collection of medical abstracts, provided by the Swiss Institute of Bioinformatics.*

MOTS-CLÉS : *Gain d'Information, Information Mutuelle, Kullback-Leibler, Machines à points supports*

KEYWORDS: *Information Gain, Mutual Information, Kullback-Leibler, Support Vector Machines*

.

## 1. Overview

Modern textual information access applications involve large text collections with large vocabularies (thousands to tens of thousands of words). In order to describe documents, Machine Learning techniques tend to rely on features defined from the words contained in the documents. In that context, feature selection methods have two main applications :

1) Select a relatively small subset of "relevant" features on which the models are learned. This has advantages in terms of training speed as well as performance, for those models that crucially depend on the dimensionality of the feature space (cf. *curse of dimensionality*).

2) Understand and explain the model decisions. This is especially important for statistical models that are often considered as "black-box" models. It allows to select the features that are most/least relevant to the decision.

In this contribution, we review a number of approaches to feature selection, divided in two broad classes. Some are corpus-based : they use only the corpus to identify the relevance of each feature. These approaches would be the preferred implementation for case 1 above, as they may easily be used prior to estimating any model. They are introduced in section 2.

In case 2, it seems that different models addressing the same decision problem may rely on different features to do so. As a consequence, we may benefit from defining a model-based relevance for each feature. Approaches to do this for different statistical models are introduced in section 3.

We are interested in the comparison between both approaches in an Information Retrieval context, and more precisely in text categorisation. Typical questions that we want to answer are : are feature that are important to the decision for a model necessarily identified as important by a simple corpus-based, model-agnostic method ? The model may implement subtle influences that the crude corpus-based approaches can not identify. Reciprocally, is corpus-based feature extraction "selective" enough, ie does it over-generate compared to model-based selection ? Clearly, there may be some model bias here : a specific model may not be able to leverage all features that appear important based on corpus calculations. These points are investigated in section 4.

It is important to note that we mainly focus here on model-based feature selection and its relation to corpus-based feature selection (aka feature selection). Our work thus differs from previous ones (e.g. [JOH 94, YAN 97, MLA 98a, MLA 98b, FOR 03]) which addressed the problem of identifying feature selection metrics appropriate for text classification or infomation retrieval. Among such works, the empirical study presented in [FOR 03] reviews a large number of potential metrics, with different characteristics and impact on performance. This study confirms the importance of the information gain as a feature selection metric, as well as introduces a new metric, the so-called bi-normal separation. Even though this latter measure outperforms the information gain in different situations, it does not generalize, at least directly, to multi-class problems. For this reason, we focus here on the information gain as the metric for (corpus-based) feature selection.

## 2. Corpus-based feature selection

We first consider methods that estimate the relevance of a feature based on the corpus alone. We assume that we have a document collection with associated category information : $\left\{(d^{(i)}, c^{(i)})\right\}_{i=1\ldots N}$, where $d^{(i)}$ is the $i$-th document and $c^{(i)}$ the category (or categories) it belongs to. Each document is represented by a set of features, which we assume may be represented by a vector $x^{(i)} = \left[{x_w}^{(i)}\right]$. Typically, ${x_w}^{(i)}$ may be the frequency of word $w$ in document $d^{(i)}$.

One way to assess the importance of a feature for a categorisation task is to estimate how much information the knowledge of the feature brings to this task. This is captured by the Mutual Information between two random variables [COL 93] : one representing the feature information, one representing the category information.

There are at least two definition of the mutual information between a feature $F$ and a category $C$ in the Information Retrieval literature. One corresponds to the Information Gain (IG, cf. [YAN 97]) or to the Likelihood Ratio [DUN 93] :

$$IG(F, C) = \sum_{f,c} p(f, c) \log \frac{p(f, c)}{p(f)p(c)} \qquad [1]$$

where the sum is over all possible values $(f, c)$ of $F$ and $C$. The other definition is a degenerate version which ignores the joint probability $p(f, c)$ in front of the $\log$. The degenerate version tends to over-estimate the importance of rare features, and we will therefore only consider the Information Gain above (eq. 1) when we talk about the mutual information.

For the category variable, we naturally use a discrete variable indicating the category, $c \in \{1, \ldots K\}$.[1] For the feature variable, the simplest solution, such as implemented by [YAN 97] is to consider a binary variable indicating the presence/absence of a feature in a document, $F \in \{0, 1\}$. This however may create problems for features for which the frequency rather than the presence is important – typically the case for documents that result from a boolean query : all terms from the query are in all documents, and have therefore no relevance according to the binary IG. A more realistic situation is to consider that the information a feature gives on a document is related to its frequency, $f \in I\!N$. This may help identify very common word as shown in figure 1. In the top right corner, we see many important terms identified by both versions of IG. The apparent diagonal corresponds to the many terms for which the IG is identical in both versions. We also identify a number of terms for which the frequency, more than

[1] If a given document may be assigned to multiple categories (multi-label), it may actually be more convenient to consider $K$ different binary variables.

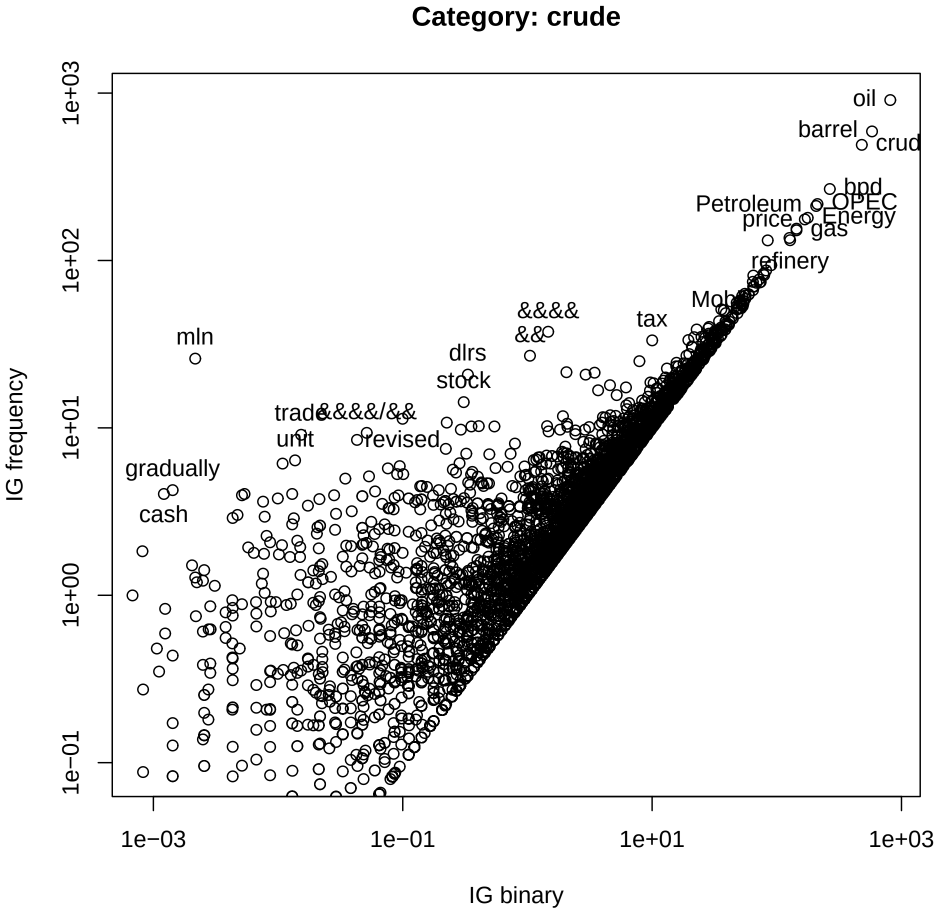

**Figure 1.** *Information Gain for terms in the Reuters collection, for categorising in the 'crude' category. Comparison between the IG calculated on the basis of the presence/absence of a term and the IG calculated on the basis of the frequency.*

the presence is important : this is the case for 'mln' and 'dlrs' (respectively 'million' and 'dollars'). As the collection contains only newswire stories from the economic domain, the presence of these terms alone is not very discriminative. However it seems that their frequency is important to discriminate documents related to oil.

## 3. Model-based feature selection

Let us now assume that we have a categorisation model. What features are important to this model when it makes a categorisation decision ? We review how we can answer this question for several models.

**Linear models :** For linear models, the score is given by $s(c,d) = \mathbf{w}^c.\mathbf{x} = \sum w_f^c x_f$, ie a weighted linear combination of the feature values $x_f$, with $\mathbf{w}^c$ the weight vector for class $c$. The importance of a feature $f$ for categorisation in class $c$ is directly related to its weight $w_f^c$. In order to take the magnitude of this feature into account and remain scale invariant, $w_f^c$ is multiplied by the standard deviation of $x_f$, $\sigma_f$. This yields the scaled regression coefficient

$$\widehat{w}_f^c = w_f^c \sigma_f \qquad [2]$$

If all features are independent, it is easy to show that $\sum_f \widehat{w}_f^c = \text{Var}\,(s(c,d))$, ie the scaled regression coefficient provides a natural decomposition of the variance of the score. Note that the decision function for Support Vector Machines (SVM) with linear kernels corresponds to the one given above. Hence, equation 2 defines a way to estimate the features supporting the categorisation decision of SVM.

**Naive Bayes :** The score is the class posterior $s(c,d) = P(c|d) \propto P(c)P(d|c)$. The “Naive Bayes” assumption is that all features are independent : $P(d|c) \propto \prod_f P(f|c)^{x_f}$, with $x_f$ the frequency of feature $f$ is document $d$. Assuming binary classification, the maximum posterior decision may be implemented by looking at the sign of the log-ratio $\ln\,(P(+|d)/P(-|d))$. All proportionality factors are independent of $c$ and disappear in the ratio, yielding :

$$\ln \frac{P(+|d)}{P(-|d)} = \sum_f x_f \ln \frac{P(f|+)}{P(f|-)} + \ln \frac{P(+)}{P(-)} \qquad [3]$$

In terms of feature values, this is a linear model. The constant $\ln \frac{P(+)}{P(-)}$ specifically models a possible bias in favour of one of the classes, while the linear coefficients are given by the $\ln \frac{P(w|+)}{P(w|-)}$. In terms of feature selection, this is identical to the linear model treated above.

**Probabilistic Latent Categoriser :** PLC [GAU 02] is a probabilistic model of the co-occurrence of features $f$ and documents $d$ : $P(d,f) = \sum_c P(c)P(d|c)P(f|c)$. The influence of a feature $f$ in a class $c$ is reflected by the probability $P(f|c)$. For example, if $P(f|+) \gg P(f|-)$ then the presence of this feature in a document will be a strong indicator that it is relevant. The difference between two probability distributions over the same event space may be evaluated using the symmetrised Kullback-Leibler divergence :

$$KL(P(f|+),P(f|-)) = \sum_f \underbrace{(P(f|+) - P(f|-)) \ln \left(\frac{P(f|+)}{P(f|-)}\right)}_{\epsilon_f} \qquad [4]$$

The divergence is zero iff both distributions are equal. Equation 4 provides a natural additive decomposition of the divergence in feature-dependent terms $\epsilon_f$. This approach has been used in [DOB 03a]. It may be extended to multiple classes by considering the KL divergence between each class-conditional feature distribution and their average $P(f)$.

We are now going to compare the different feature selection mechanisms we have reviewed.

## 4. Experimental comparison

We conducted a qualitative and quantitative comparison of the different measures presented above on two collections : a collection of news articles (namely Reuters), and a collection of medical abstracts, which we refer to as the Swiss-Prot collection. We present here the results we obtained on the the Swiss-Prot collection. The first results obtained on the Reuters collection are similar, and will be presented in the final version of the paper.

The Swiss-Prot collection contains 2188 titles and abstracts from 32 different genes. These documents were selected from PubMed using a query that takes the general form : *<gene name> AND ((mutations OR mutation) OR (variants OR variant) OR (polymorphisms OR polymorphism))*, and were then reviewed by medical annotators who assigned each document to one of the categories : *Good*, ie relevant for medical annotation, *Bad*, ie irrelevant for this annotation, or *Unclear*, when the title and abstract do not contain enough information to make a decision. Overall, 14% of our collection was assigned to the *Good* class, 70% to the *Bad* class, and 16% to the *Unclear* one (see [DOB 03b] for more details on this collection).

In order to study the correlation between the different measures, we make use of standard coefficient, namely the linear correlation coefficient, and the Spearman correlation coefficient which, based solely on a rank comparison, pressents the advantage of being independent of the scale used.

### *Binary vs. frequency-based information gain*

The comparison between the binary and frequency based Information Gain (IG) shows a large similarity between the twos on the Swiss-Prot collection (see figure 2). Indeed, the correlation on the logarithmic scale is 0.652, while the Spearman correlation coefficient reaches 0.715. When comparing the top 50 terms of both lists, one finds that they are mostly similar, with only 7 terms differing. Moreover, most of these terms are still high ranked in the other list. Two noticeable exceptions are : *gene_req* (a string used to denote all occurrences of the gene name used in the original queries) and *DESOXYRIBONUCLEIC_ACID*. The former has a lower rank with the frequency-based IG than with the binary IG. In this case, the frequentist version of IG is able to capture the fact that an abstract in which the gene name of the original query often occurs is more likely to contain interesting information about the gene under focus than an abstract in which the gene name is barely mentioned, whereas the binary IG is unable to account for such a distinction. The latter term (*DESOXY-RIBONUCLEIC_ACID*) takes rank 256 with the binary IG, and 43 with the frequency based IG. It corresponds to a generic token used to replace various DNA spellings and seems to be a negative selector. Indeed, articles in which this word occurs often deal with cloning or genetics and are irrelevant in our context. Our findings show that the more frequent this word is in a document, the less relevant the document is, a fact that the frequency based version of IG is able to capture, but not the binary one. Because of this property, we focus, in the remainder, on the frequency-based IG.

### *Document frequency vs. information gain*

As noted by [YAN 97], the Information Gain (IG) is linked to the $\chi^2$ statistic, which is another way to measure the (in)dependance of two variables from a contingency table. The IG is also similar to the Likelihood ratio statistic, as presented, eg by [DUN 93]. Finaly, [YAN 97] also noted that, surprisingly, the Information Gain seemed somewhat correlated with the Document Frequency (DF) on the Reuters collection. This counters the usual Information Retrieval intuition that terms with large DF are less informative. However, we find that the effect reported in [YAN 97] is not compelling. First, it is not necessarily observed on other collection such as the Swiss-Prot collection, as illustrated in figure 3. Second, even on the Reuters corpus, the correlation appears only in the log domain, and the variability in DF for large values of IG is so large that there is essentially no useful relationship between the twos on a normal scale. Finally, we believe that this effect may be artefactual : as mentioned earlier, the Information Gain does in a way “down-play” the importance of low-frequency terms, which may “tilt” a perfectly uncorrelated relationship to display some apparent correlation.

On the Swiss-Prot corpus, 39 terms, among the union of the first 50 ones for each measure, differ between IG and DF, and while most of the terms selected by IG are still highly ranked with DF (in the top 300) the reverse is not true. This is confirmed by the Spearman correlation coefficient, which amounts to only 0.578.

### *Information Gain vs. Kullback-Leibler*

The comparison between the Information Gain (IG) and the Kullback-Leibler divergence (KL) (see equation 4) on the Swiss-Prot collection shows that while the two differ (18 different terms among the union of the first 50 ones for each measure), most of the terms deemed important by one measure are also deemed important by the other. The linear correlation between the twos on a log-scale reaches 0.726, whereas the Spearman correlation coefficient amounts to 0.686. This correlation is shown on figure 4.

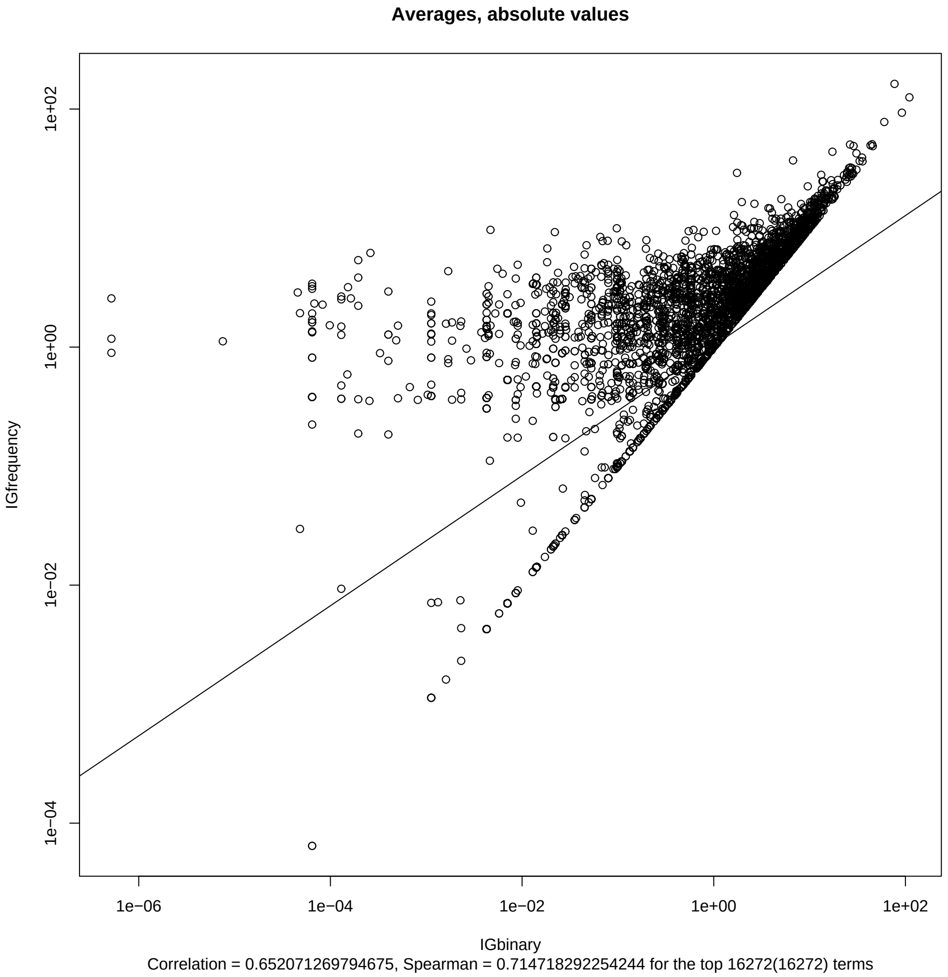

**Figure 2.** *Binary vs. frequency based Information Gain in the Swiss-Prot collection.*

Nevertheless, two terms behave rather differently according to the measure used : *gene*, ranking 28 with IG and 644 with KL, and *XERODERMA_PIGMENTOSUM*, ranking 42 with IG and 1930 with KL. The latter is a generic name for two different genes of the same family in several species. Our detailed examination of it showed that it was not particularly correlated with irrelevance since its distribution in the different classes mirror the proportion of the classes. In this case, IG seems to be overly sensitive to the tail of the distribution, which mainly consists of low counts.

The former is a good example where viewing features in isolation (ie indenpendently of each other) may lead to a wrong judgement of the importance of each fea-

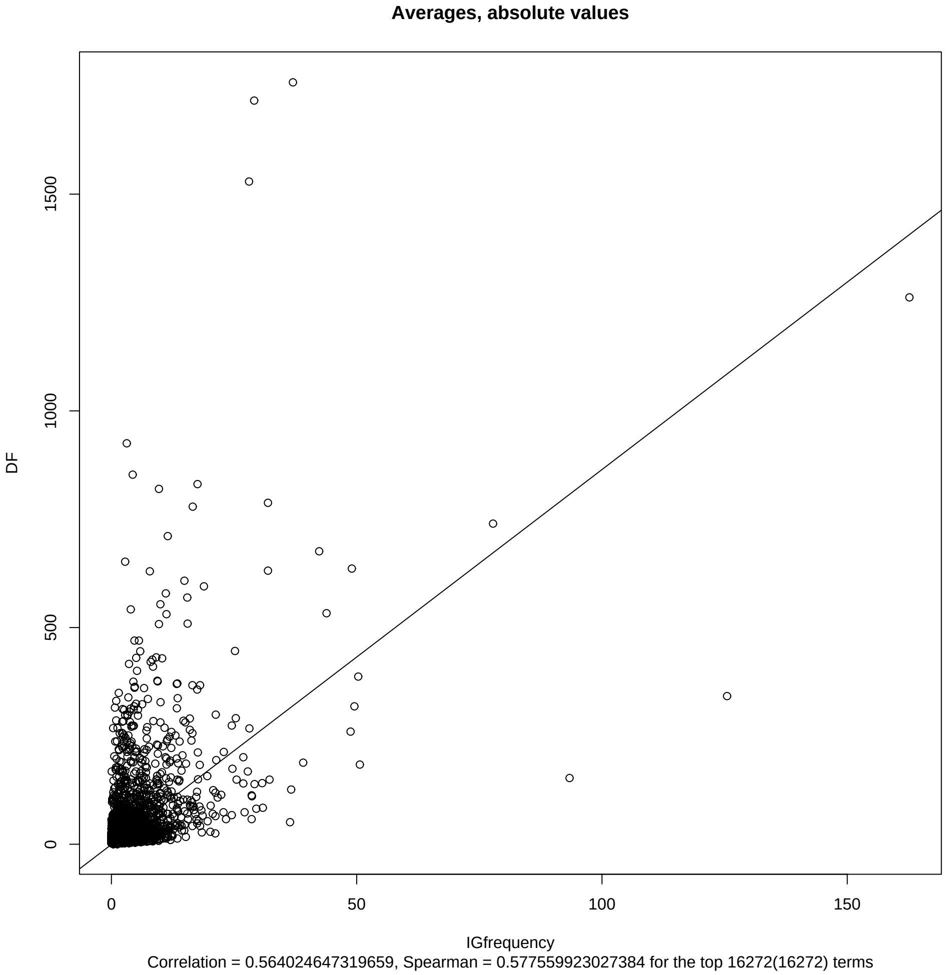

**Figure 3.** *Information Gain versus Document Frequency in the Swiss-Prot collection.*

ture. Typically, the word *gene* occurs in combination with a nominal gene name, as in <*gene_name*> *gene*, or *gene of* <*gene_name*>. In such cases, the selective term is <*gene_name*> (eg the name of the queried gene), but not the word *gene* itself, a fact that KL seems more amenable to capture than IG.

Another interesting observation arises from the examination of the frequency distribution of this term. It seems that the word *gene* plays a selective role for the *Bad* class when it is either absent (0 occurrence) or very frequent (more than 10 occurrences). Table 1 shows how documents fall within each of the two classes *Bad* (C1)

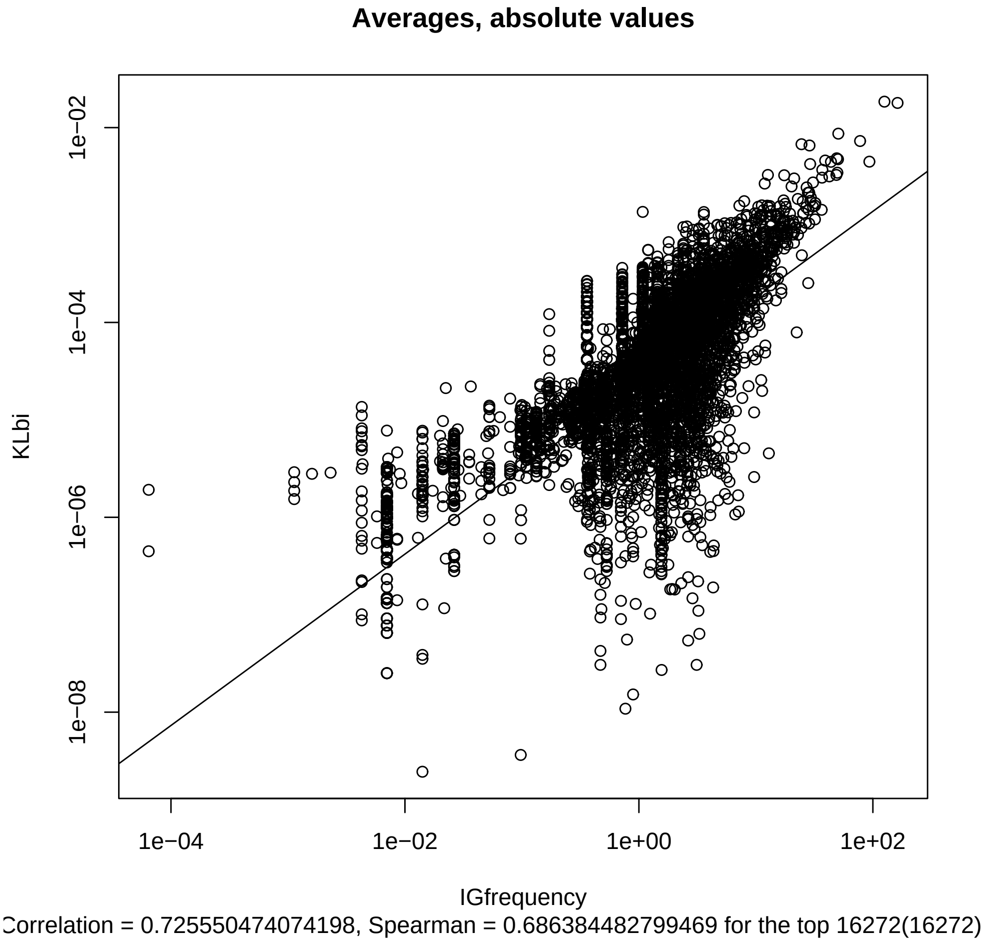

**Figure 4.** *Information Gain versus Kullback-Leibler divergence in the Swiss-Prot collection.*

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 16 | 17 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C1 | 510 | 292 | 234 | 163 | 133 | 84 | 37 | 35 | 18 | 5 | 7 | 5 | | 2 | 2 | | 1 |
| C2 | 149 | 162 | 126 | 83 | 37 | 18 | 14 | 1 | 2 | | | 1 | | | | 1 | |

**Tableau 1.** *Distribution of documents according to the frequency of gene.*

and *Good or Unclear* (C2) depending on how frequent the word gene is in these documents.

While the fact that no occurrence of *gene* in a document is a good indicator that the document is not relevant seems logical (an abstract without this word is unlikely to deal with mutations or polymorphisms), the fact that a high frequency of this word entails irrelevance is less obvious. By manually examining the abstracts in which this word occurs, several hypotheses can be formulated :

1) The abstract contains the word in its plural form (*genes*) and is of review/tutorial type. It thus deals with several gene families, rather than a single gene.

2) The abstract covers a number of different genes, but does not contain indepth information about a particular one.

Whatever the actual explanation, the dependence of the importance of this feature on its frequency is nicely captured by IG, but missed by KL, which averages over the number of occurrences of the feature in the different documents it occurs in. On the other hand, KL is able to rely on several gene names and abbreviations that IG fails to capture.

### *Information gain vs. scaled regression coefficient*

The comparison between the Information Gain (IG) and the scaled regression coeeficient applied on linear kernel Support Vector Machines (SVM) shows, on the Swiss-Prot collection, that the two measures are not correlated (see figure 5). The correlation coefficient is 0.433, while the Spearman correlation coefficient only reaches 0.536.

Among the union of the first 50 terms for both measures, 34 differ, and many terms which are deemed important by IG are only marginally used in SVM. We attribute this to the fact that the scaled regression coefficient for SVM selects terms present in the support vectors, and are thus close to the frontier between the different classes. In other words, whereas IG tries to select the most important, ie central, terms for each class, the scaled regression coefficient for SVM will tend to select the less central ones.

## 5. Discussion

The above comparison leads us to postulate several facts concerning the different measures we have investigated. First, concerning the Information Gain, our evaluation shows that even though the binary and frequentist versions are highly correlated, the frequentist one is more appropriate as a feature selection method for the model we considered (Probabilistic Latent Categoriser and Support Vector Machines), since its correlation with other measures is higher than the one of the binary version (table 2 summarizes the correlations between the different measuress). This comes as no surprise since these models rely on the frequency of each feature, and not on their mere presence/absence. An interesting point we can note wrt the frequentist version of the information gain is its ability to spot those features of which the importance lies in their frequency distribution, and cannot be captured through a single, summarising statistics (as for eg the word *gene* discussed above).

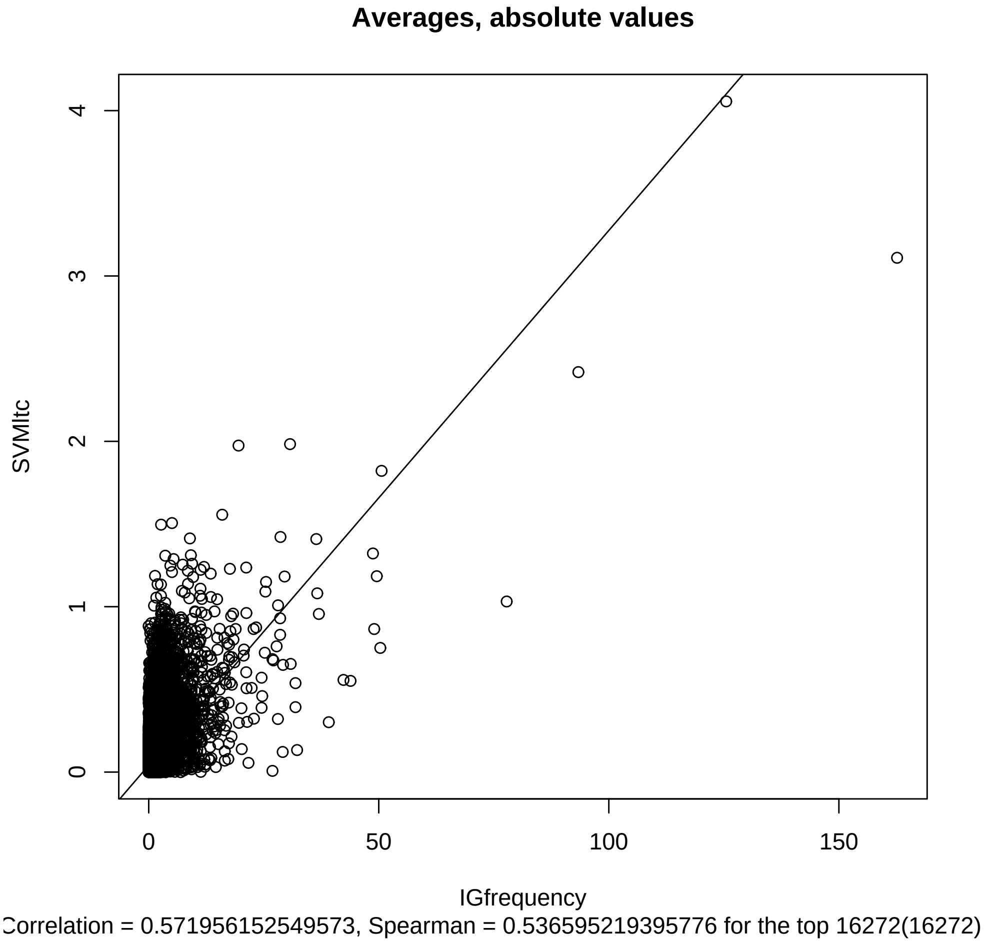

**Figure 5.** *Information Gain versus scaled regression coefficient for linear Support Vector Machines in the Swiss-Prot collection.*

Another point we deem important to mention is the fact that we failed to confirm the correlation between the document frequency and the information gain reported in [YAN 97]. In particular, the Spearman correlation coefficient for those measures amounts only to 0.578 on the Swiss-Prot collection, which does not reveal a correlation between the two measures. This absence of correlation is well in line with the common information retrieval assumption that terms with large document frequency are less informative.

The comparison between corpus-based and model-based feature selection we performed call for several comments. First, the two approaches have somewhat different

| | *IGf\IGb* | *IGf\DF* | *IGf\KL* | *IGf\Scal. reg.* | *DF\KL* | *DF\Scal. reg.* | *IGb\DF* | *IGb\KL* | *IGb\Scal. reg.* | *KL\Scal. reg.* |
|---|---|---|---|---|---|---|---|---|---|---|
| *log-log* | 0.652 | 0.484 | 0.726 | 0.433 | 0.430 | 0.677 | 0.089 | 0.656 | 0.187 | 0.432 |
| *Spearman* | 0.715 | 0.578 | 0.686 | 0.536 | 0.411 | 0.881 | 0.232 | 0.679 | 0.250 | 0.365 |

**Tableau 2.** *Linear and Spearman corrrelation coefficients for the different measures. Note that the scale for the scaled regression coefficient is always linear.*

goals, the former aiming at selecting a small subset of features prior to the construction of a categoriser, the latter primarily aiming at explaining the categorisation decisions. This difference is particularly highlithed when comparing the information gain with the scaled regression coefficient for linear SVM. In such a case, the two measures rely on two different notions of the "importance of a feature" : a feature is important if it is representative of a category for the information gain, whereas it is important it is close to the decision boundary for the scaled regression coefficient. We plan to investigate new measures in line with the first definition of importance for linear models such as SVM.

The comparison between the information gain and the Kullback-Leibler divergence shows a good correlation between the twos, indicating that the information gain does a proper job at selecting those features at the core of probabilistic models such as PLC [GAU 02], or, even though not apparent in this study, Naive Bayes. However, because of their difference, the information gain cannot be used in lieu of the Kullback-Leibler divergence when it comes to explaining the categorisation decision of the models. Furthermore, the detailed study we conducted on these two measures reveals some advantages of each approach : the information gain is able to capture those features the importance of which depends on the frequency range, a property that the Kullback-Leibler divergence does not display ; on the other hand, PLC coupled with the Kullback-Leibler divergence seems able to deal more accurately with redundant features.

Lastly, we want to mention an interesting side effect of feature selection, which is to allow one to assess and uncover the shortcomings of the various preprocessing steps. In our case, synonym terms which were not properly normalised, multiword terms which were not properly delimited show up high in the list of say KL, when those terms correspond to gene names important for the categorisation.

## 6. Conclusion

We have presented in this article different approaches to feature selection, that address different problems : selecting only a subset of features prior to the development of categorisers, and explaining the decisions made by different categorisers. We have proposed several measures to this end, for models ranging from Naive Bayes to Probabilistic Latent Categorisers and Support Vector Machines. Furthermore, we have conducted a detailed comparison between the different approaches, showing the advantages of each of them. This study also led us to question a claim made in [YAN 97] on the correlation between two widely-used IR measures : the information gain and the document frequency.

## 7. Bibliographie

[COL 93] Colin B., « Information et analyse des données », *Pub. Inst. Stat. Univ. Paris*, vol. XXXVII, n$^{o}$ 3–4, 1993, p. 43–60.

[DOB 03a] Dobrokhotov P., Goutte C., Veuthey A.-L., Gaussier E., « Combining NLP and Probabilistic Categorisation for Document and Term Selection for Swiss-Prot Medical Annotation », *Proceedings of ISMB-03, Bioinformatics*, 2003, p. 191–194.

[DOB 03b] Dobrokhotov P., Goutte C., Veuthey A.-L., Gaussier E., « A Probabilistic Information Retrieval Approach to Medical Annotation in SWISS-PROT », *Proceedings of MIE-03, Medical Informatics Europe*, 2003.

[DUN 93] Dunning T., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, vol. 19, n$^{o}$ 1, 1993, p. 61–74.

[FOR 03] Forman G., « An Extensive Empirical Study of Feature Selection Metrics for Text Classification », *Journal of Machine Learning Research*, vol. 3, 2003, p. 1289–1305.

[GAU 02] Gaussier E., Goutte C., Popat K., Chen F., « A Hierarchical Model for Clustering and Categorising Documents », Crestani F., Girolami M., van Rijsbergen C. J., Eds., *Advances in Information Retrieval—Proceedings of the 24th BCS-IRSG European Colloquium on IR Research*, vol. 2291 de *Lecture Notes in Computer Science*, Springer, 2002, p. 229–247.

[JOH 94] John G., Kohavi R., Pfleger K., « Irrelevant Features and the Subset Selection Problem », *Proceedings of ICML-94, 11th International Conference on Machine Learning*, 1994, p. 121–129.

[MLA 98a] Mladenić D., « Feature Subset Selection in Text Learning », *Proceedings of ECML-98, European Conference on Machine Learning*, 1998.

[MLA 98b] Mladenić D., Grobelnik M., « Feature Selection for Classification based on Text Hierarchy », *Proceedings of CONALD-98, Conference on Automated Learning and Discovery*, 1998.

[YAN 97] Yang Y., Pedersen J. O., « A Comparative Study on Feature Selection in Text Categorization », *Proceedings of ICML-97, 14th International Conference on Machine Learning*, 1997, p. 412–420.