Topic 1: Look, Mom, no hands!

Assignment 2
Topics in Algorithmic Data Analysis SS'17
Submitted to: Dr. Jilles Vreeken
Due by: June 01, 2017 before 10.00 hours

Name: Harshita Jhavar
Matriculation Number: 2566267
eMail: s8hajhav@stud.uni-saarland.de

Introduction

We can never be sure if the assumed parameters for the data mining algorithms are the best for the given data. Incorrect parameters can lead to complex computations. In worst, this can result in failure to find the underlying true patterns which defeats the whole purpose of data mining. They can also lead to a bias towards detecting patterns which might not be existing in reality. Thus, 'Parameter free methods' based on the concept of Minimum Description Length(MDL) is hot favourite for data mining. However, while using these methods, one has to keep in mind their vulnerabilities which are defined by the hidden assumptions underlying their structures. In this report, I will discuss about the heart and soul of these parameter free methods based on MDL for data mining and highlight my evaluatory understanding for the same. The road map for this report will be to discuss about the advantages of the parameter free techniques, followed by assumptions and problems for the MDL based parameter free techniques and finally, ending the report with conclusion.

What makes Parameter Free Methods Advantageous

I agree that the parameter free methods have their own set of advantages. The experimentor now doesn't have to use his/her assumptions and beliefs to intuitively decide value of any parameter. Again, unlike parametrized algorithms, there are parameter free methods which do not demand any special structure of the input data. This rules out many pre-processing computations and makes our experiments easier. On its positive side, I agree that now it is with these methods that it can get easier to reproduce the published experimental results, independent of any dependecies of hardware or values of parameter used and extend the line of research as done in Keogh[3]. I can argue the requirement of the parameter free algorithms with the parallel existence of parameter self tuning algorithms, for example, back propogation but ironically, one needs to set parameters for these parameter tuning algorithm first which is not the case for parameter free methods. It will be more profitable if we use compression as the atomic idea behind the design of such parameter-free algorithms written for tasks like, summarising raw data for further analysis as done in Mampaey and Vreeken[2] using clustering or other data mining problems like 'anomaly detection', 'classification' as done in Keogh[3]. However, I am still dubious and is taking these algorithms as 'parameter free' with a pinch of salt as discussed in coming sections.

Parameter Free Methods based on MDL

Parameter free methods based on MDL have compression as the atomic idea behind their structure. Compression algorithms are space and time efficient as experimentally proven in Keogh[3] which makes this approach attractive for humongous data sets. MDL restricts the set of allowed codes in such a way that it becomes possible to find the shortest code length of the input data, relative to the allowed codes which is shown in Grnwald[1]. I understand that this is how MDL overcomes the problem of Kolmogorov complexity's uncomputability as it is impossible to write an algorithm which outputs the shortest program that produces the data. Again, in large datasets, the language used for encoding might not matter but I really support the argument that for smaller datasets, the constants diregarded in Kolmogorov complexity computations definitely have an influence. So, is parameter free methods based on MDL the best thing since sliced bread? No, I argue the claim for these to be the future for data mining techniques with reasons in the next section.

Wait! There are some hidden assumptions and problems here!

While the MDL based parameter free techniques seem very appealing, however, I believe that there are certain hidden assumptions:

- Firstly, it is assumed here that the sample data is a true representative of the entire possible values of the population. For example, in Mampaey and Vreeken[2], for coming up with the code table for each cluster, if a certain value v in the sample had a frequency of 0, which implies that it doesn't occur in the sample data, there is no code recorded for this value v in the code table. How will this clustering approach account for the unseen data values whose code doesn't exist in the code table. There will always be some regular sequences which will not be able to be compressed. It is simply loss of information and productivity from hard-earned dataset.

- Secondly, I do not know here on how are you deciding on what is the shortest acceptable length? What is the definition of 'short' here? The probability of a hypothesis dynamically changes as the learner knows that something can be further compressed as 'short length' does not have any standard definition and is subjective.

- Thirdly, it might be possible that with the motive of getting the most compressed representation of data, the theory lying behind the data might become intractable. While the parameter free approach gives the veto power to the data, unlike the algorithms with different parameters controlled by user; the results can still result in a false pattern detection. There might be patterns detected which are not true in general for the entire data. Thus, patterns detection from these noisy data might result in total failure. We are giving too much power to the compressed sample data assuming it to be the 'true' representation of the population which might not be the case always.

- Fourthly, in the fundamental idea of MDL: more the data is compressed, the more we have learned about the data; I find another strong assumption that we believe that the future data will behave according to the law or patterns showed by the sample data. Though, MDL principal is an almost optimal choice for the universal distribution but the universal distribution here is just a choice when we do not have any information at all about the real origin of the information.

- Fifthly, I totally argue that the atomic concept of CDM dissimilarity measure that the side containing the most unusual section will be less similar to global sequence than other half, has another assumption. I do not understand how we are ensuring the global sequence is 'global' in literal terms. How can one confirm which side is interesting when it might be possible that the side which seemed uninteresting in the beginning might actually contain some unusual patterns whose values were nullified by lot many other trivial or 'uninteresting' patterns and thus, were not detected. Above all, these algorithms do not seem to be 'parameter free' in true sense as one has to assume various thresholds to define 'what is interesting', 'language of encoding', 'threshold for minimum encoding length', 'CDM dissimilarity thresholds', correlation detection parameters assumptions on thresholds based on kolmogorov complexity etc. Thus, parameter free assumptions are not actually absolutely parameter free.

- Sixthly, I strongly believe that the parameter free techniques based on MDL for clustering, classification etc. might not be successful for all domains of data. For example: the techniques discussed in Mampaey and Vreeken[2] for clustering and Keogh[3] for classification expect to separate structure from the noise. However, there are times when the entirely possible structure can not be encoded as there is no possible formal language to encode such structures. Consider, the problem of encoding grammar for English language. One can never be sure that a particular encoding of structure explains all rules and cases of very large and versatile grammar. This is a very hard problem in Computational Linguistics and Textual Information Retrieval. I cannot extract all meaningful information by encoding say, the novel data to study fictional characters, as I do not have any formal language in which I can explain the entire novel text in terms of a grammar rules and logics. Note: Thinking about defining semantic aspects of the textual data is beyond the scope of formal language encoding currently as we are currently not sure on how the cognitive understandings work. Infact, we are currently not yet done with formalizing the entire syntactic aspect of the data ex: grammar. Formalizing semantic or any cognitive aspects behind textual data, image interpretation, datasets from computer vision, psycholinguistics etc cannot always be formally encoded as we do not know what the cognitive rules are all about. We do not always have a formal language to define all kinds of data.

MDL: a Magic Wand or just another floo powder

I agree that the MDL philosophy explained in Grnwald[1] is positive, but MDL technique should only be used for large and non-random data from noisy sources. If the model generated from MDL techniques is not much predictable, at least it is short and manageable. In MDL based techniques, there is no goal for explanation, no idea of surprise or interest in the result of the learning algorithm.

In figure 1, I have summarised my understanding from Grnwald[1]. I can conclude from these MDL techniques that even though they allow us to compare two models of different functional form, still, it might end up choosing lesser complex model with the same number of parameters even though the more complex model might fit the test data well. Again, MDL talks about

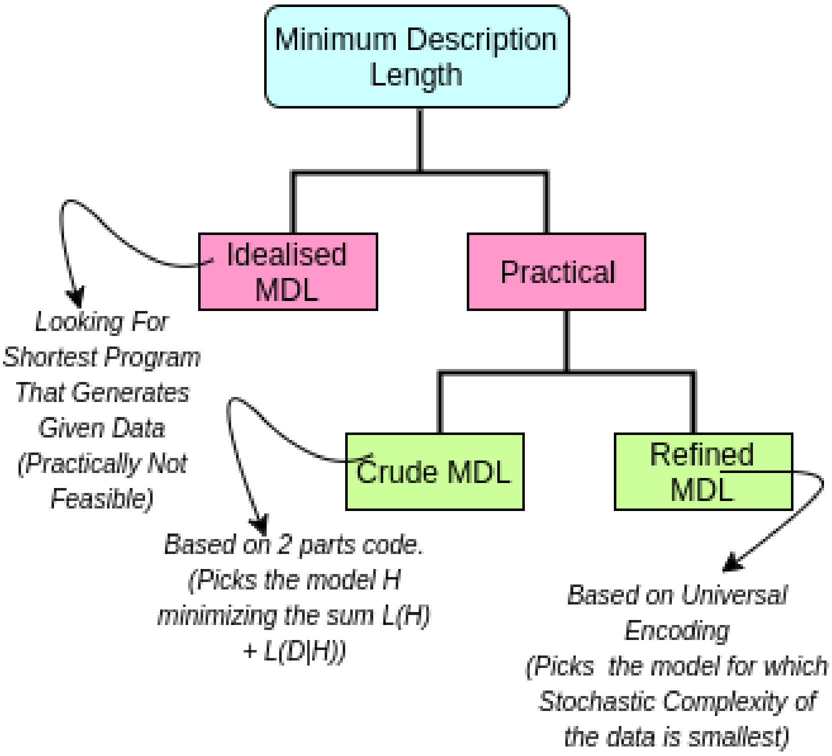
Minimum Description Length
Idealised MDL
Practical
Looking For Shortest Program That Generates Given Data (Practically Not Feasible)
Crude MDL
Refined MDL
Based on 2 parts code. (Picks the model H minimizing the sum L(H) + L(D|H))
Based on Universal Encoding (Picks the model for which Stochastic Complexity of the data is smallest)

Figure 1: *Approaches in MDL for model selection*

data but not about the assumptions and hypothesis underlying the model defined for that data. Parameter free techniques based on compression are argued as future of data mining in Keogh[3]. However, in a case when most of the data sequences are noncompressible, the MDL principle gives no knowledge at all, in general. No knowledge may be harmless but will be useless as well. Also, I cannot totally take the MDL techniques as 'parameter-free' as one has to decide the choice of the specific compressor to be used which has its own set of compression parameters. For compression techniques like CDM as discussed in Keogh[3], the input data have to be recorded with the same set of precision which might lead to more pre-processing steps than usual. I totally argue that even though we might think of converting the input data into discreet form with some kind of encoding technique, again not all data can be encoded that way. Ex: For generating knowledge graph representations of textual data, one cannot encode the entire set of possible relations in the different entities present in data. So, choice of the representation of the data is again giving more power to the experimentor rather than the data itself. Also, the inferences based on encoding to find relations between entities in knowledge graphs might be lesser now as we might lose some possible information and inferences while compressing the data. Thus, while parameter free algorithms are not totally parameter free, still, compression based techniques and MDL based techniques are good way to ensure a trade-off between goodness of fit and lesser complexity of the models. I would rather stick with the term 'Parameter Light' instead of 'Parameter Free' techniques based on MDL.

In conclusion, the MDL principle works well in those environments where the bias does not allow extensional descriptions or where the data are huge and from statistical or imperfect sources. But, when faced with a concrete learning problem or in scientific discovery, alternatives could be to tune length, computational time, intensionality and informativeness of descriptions according to the expectation we have about the source of knowledge.

References

[1] Grnwald, P. *Minimum Description Length Tutorial (shortened version).* In Advances in Minimum Description Length, MIT Press, 2005.

[2] Mampaey, M. and Vreeken, J. *Summarising Data by Clustering Items.* In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Barcelona, Spain, pages 321-336, Springer, 2010.

[3] Keogh, E., Lonardi, S. and Ratanamahatana, C.A. *Towards parameter-free data mining.* In Proceedings of the 10th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Seattle, WA, pages 206-215, 2004.

[4] Hernndez-Orallo, J. and Garca-Varea, I. *Explanatory and Creative Alternatives to the MDL priciple* Foundations of Science (2000) 5: 185. doi:10.1023/A:1011350914776