**Rule Based Systems and Networks for Knowledge Discovery in Big Data**

|  |  |
| --- | --- |
| **Alexander Gegov**  *University of Portsmouth School of Computing*  *Buckingham Building, Lion Terrace*  *Portsmouth PO1 3HE, UK* | alexander.gegov@port.ac.uk |
|  |  |
| **David Sanders**  *University of Portsmouth School of Engineering*  *Anglesea Building, Anglesea Road*  *Portsmouth PO1 3DJ, UK* | david.sanders@port.ac.uk |

**Abstract**

This paper discusses rule based systems and networks in computational intelligence and their use for machine learning in the context of knowledge discovery in big data. Rule based systems and networks are seen as a special type of expert systems that can be built by using expert knowledge or learning from real data. From this point of view, the design of rule based systems and networks can be divided into expert and data based design. In the present big data era, the latter approach that is based mainly on machine learning has become quite popular.

**Keywords**: rule based approaches, machine learning algorithms, big data processing

1. **Rule Based Approaches in Machine Learning**

A rule based system or network is a special type of expert system. It usually consists of a set or sets of if-then rules that can serve many purposes such as decision support or predictive decision making (Gegov, 2011). One of the main challenges in this area is the design of such systems which could be based on both expert knowledge and data. Thus, the design techniques can be divided into two categories: expert and data based. The former follows a traditional engineering approach while the later follows a machine learning approach. For both approaches, the design of the rule based system or network could be used for tasks such as classification, regression and association.

The data based approach is usually preferred to the expert based approach (Liu et al., 2016). This is because the expert based approach has some limitations which can usually be overcome by using the data based approach. For example, expert knowledge may be incomplete or inaccurate; some of experts’ points of view may be biased; engineers may misunderstand requirements or have technical designs with defects. When problems with high complexity are dealt with, it is difficult for both domain experts and engineers to have all possible cases considered or to have perfect technical designs. Once a failure arises with an expert system, experts or engineers may have to find the problem and fix it by reanalysing or redesigning. However, the real world has been full of big data. Some previously unknown information or knowledge could be discovered from data. Data could potentially be used as supporting evidence to reflect some useful and important pattern by using modelling techniques. More importantly, the model could be revised automatically as a database is updated in real time when data based modelling technique is used. Therefore, the data based approach would be more suitable than the expert based approach for construction of complex rule based systems and networks.

Machine learning is a branch of computational intelligence and involves two stages: training and testing. Training aims to learn something from known properties by using learning algorithms and testing aims to make predictions on unknown properties by using the knowledge learned in the training stage. From this point of view, training and testing are also known as learning and prediction respectively. In practice, a machine learning task aims to build a model that is further used to make predictions by adopting learning algorithms. This task is usually referred to as predictive modelling. Machine learning could be divided into two types: supervised learning and unsupervised learning, in accordance with the form of learning. Supervised learning means learning with a teacher because all instances from a training set are labelled. The aim of this type of learning is to build a model by learning from labelled data and then to make predictions on other unlabelled instances with regard to the value of a predicted attribute. The predicted value of an attribute could be either discrete or continuous. Therefore, supervised learning could be involved in both classification and regression tasks for categorical prediction and numerical prediction, respectively. In contrast, unsupervised learning means learning without a teacher. This is because all instances from a training set are unlabelled. The aim of this type of learning is to find previously unknown patterns from data sets. It includes association, which aims to identify correlations between attributes, and clustering, which aims to group objects based on similarity measures.

Machine learning is also used for data mining to discover some previously unknown patterns. This task is usually referred to as knowledge discovery. From this point of view, data mining tasks also involve classification, regression, association and clustering. Both classification and regression can be used to reflect the correlation between multiple independent variables and a single dependent variable. The difference between classification and regression is that the former typically reflects the correlation in qualitative aspects, whereas the latter reflects in quantitative aspects. Association is used to reflect the correlation between multiple independent variables and multiple dependent variables in both qualitative and quantitative aspects. Clustering can be used to reflect patterns in relation to grouping of objects.

1. **Machine Learning Algorithms for Big Data**

Big data and machine learning have been treated as connected activities. In particular, the relationship between big data and machine learning is very similar to that between life experience and human learning. In this context, people learn from their past experiences to deal with unfamiliar matters. Similarly, machines learn from big data to resolve newly challenging issues.

In the context of machine learning, learning algorithms are typically evaluated in terms of accuracy, efficiency and interpretability. These three dimensions can be strongly related to veracity, volume and variety respectively.

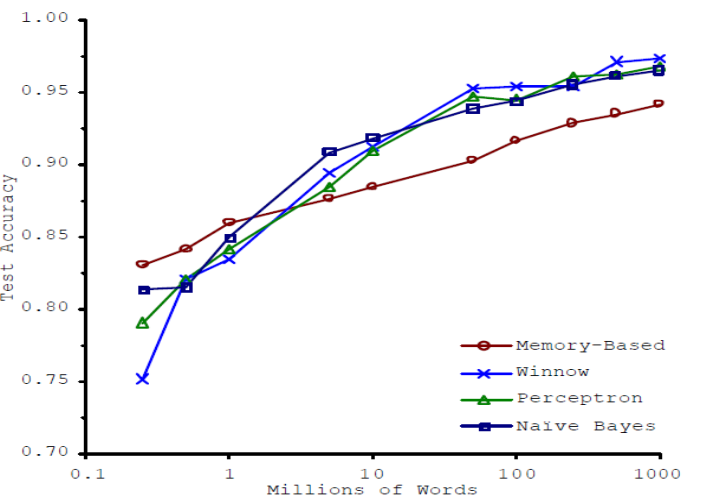
Veracity reflects the degree to which data can be trusted as mentioned above. In practice, the degree of trust is strongly dependent on the information/knowledge discovered from the data. This is because data usually needs to be transformed to information/knowledge prior to its actual use. Therefore, evaluation of the degree of trust for particular data can be done through estimation of the accuracy of the models built on the data.

Figure 1: Improvement of words disambiguation by learning from big data

Volume reflects the size of data. In machine learning and similar areas, the data size can be measured by the product of data dimensionality and sample size, i.e. the higher the data dimensionality/sample size the larger the data. Therefore, evaluation of the volume for particular data can be done through checking the data dimensionality and its sample size.

Variety reflects the format of data. In machine learning, the data format can be related to data types and presentation. In particular, data types include nominal, ordinary, string, Boolean, integer and real etc. More details on data types can be seen in (Tan et al., 2006). In practice, data types are simply divided into two categories: discrete and continuous, in machine learning tasks. On the other hand, data can be represented in different forms such as text, graph and tables. This strongly relates to model representation since model is a form of information/knowledge that is transformed from data and can be represented in different forms.

Both data types and representation can impact on interpretability due to human characteristics. In particular, with regard to data types, people in some areas such as engineering and mathematics mostly deal with quantitative data so they normally prefer to see data with continuous values that reflect quantitative aspects. With regard to data representation, these people would prefer to see data in the form of diagrams or mathematical notations. In contrast, people in other areas such as humanities and social science mostly deal with qualitative data so they normally prefer to see data with discrete values that reflect qualitative aspects. With regard to data representation, these people would prefer to see data in the form of text.



1. **Performance Evaluation Indicators in Machine Learning**

As already discussed, machine learning algorithms are usually evaluated against accuracy, efficiency and interpretability. The presence of big data has deeply affected machine learning tasks in the three aspects mentioned above.

In terms of accuracy, overfitting of training data can be significantly reduced in general as the size of data is greatly increased. There is evidence reported that learning from large training sets can significantly improve the performance in predictive modelling. The evidence is illustrated in Figure 1 from (Banko et al., 2001) which shows that the complex problem of learning on automated word disambiguation would keep improving after the size of training data is beyond billion in words. The improvement in learning performance is due to the fact that the increase in data size can usually improve the completeness of the pattern covered. In other words, small data may be supposed to cover only a small part of pattern in a hypothesis space. Therefore, overfitting of training data is likely to result in the case that a learning algorithm may build a model that performs greatly on training data but poorly on test data. This case occurs especially while the training data covers the pattern that is slightly different from that in test data. When the size of data is increased, the training data is likely to cover the pattern that is more similar to with that exists in test data.

On the other hand, the increase in data size may also increase the chance to have noise and coincidental pattern present in the data. This is due to the fact that the biased improvement in the quantity of data may result in the loss of quality. In addition, large training data may cover some patterns which occur in very low frequencies. This could mean that the pattern covered by the training data is purely coincidental rather than scientifically confident.

The above issues regarding accuracy can be solved through scaling up algorithms or scaling down data. The former way is to reduce the bias on algorithms side. In particular, the algorithms can be designed to be more robust to noise and avoid being confused by coincidental patterns. In the context of rule learning, the reduction of bias can be achieved through direct advancement of rule generation methods or employment of rule simplification algorithms. The latter way is to reduce the variance on data side. In particular, data can be pre-processed through removal of irrelevant attributes by feature selection techniques or merge of redundant attributes by feature extraction techniques. In addition, data can also be resampled by selecting only those instances that are more representative.

In terms of efficiency, the increase in the size of data usually increases the computational costs in both training and testing stages. In the training stage, it may take much longer to build a predictive model by learning from big data. In the testing stage, the model is likely to be built to have a high level of complexity, which significantly slows down the process of predicting on unseen instances. In particular to rule based systems, big data may result in the generation of large number of complex rules.

Processing of big data needs decomposition, parallelism, modularity and recurrence. In this case, some machine learning algorithms, which are inflexible and work in black box manners, would be failed in dealing with big data. This case would immediately happen to those algorithms that are quadratically complex (O (n2)), when encountering data with millions of points (instances).

The above issues regarding efficiency can also be solved through scaling up algorithms or scaling down data. In the former way, the algorithms can be designed to have a low computational complexity in training stage and thus much less affected by the increase in the size of training data. In addition, the improvement of efficiency can also be achieved through the employment of rule simplification methods as some of such methods can stop the process of rule learning earlier. In the latter way, the data size can be reduced through dimensionality reduction and data sampling. This not only reduces the computational costs in the training stage but also results in the generation of simpler models and thus speeds up the process of predicting on unseen instances in the testing stage.

In terms of interpretability, the increase in the size of data usually decreases the interpretability. The latter can be affected by the size of training data in terms of model complexity. In the context of rule based systems and networks, big data may result in the generation of large number of complex rules or sets of rules that would make it difficult for people to read and understand.

The above issues regarding interpretability can also be solved through scaling up algorithms or scaling down data. In the former way, the algorithms can be designed to be robust to noise and irrelevant/redundant attributes. In particular, the presence of noise and irrelevant/redundant attributes cannot make these algorithms learn irrelevant patterns. In the context of rule learning, rule generation algorithms may decide to skip some attributes/attribute-value pairs for generation of decision trees or if-then rules due to irrelevance. In addition, the employment of rule simplification methods also helps improve the interpretability since the employment usually results in the generation of smaller number of simpler rules. In the latter way, the data size is reduced through dimensionality reduction and data sampling as mentioned above. In particular, the reduction of dimensionality decreases the maximum number of rule terms of each single rule. The data sampling also reduces the maximum number of rules. In this way, the interpretability can be improved if the dimensionality reduction and data sampling are done effectively.

**References**

M. Banko and E. Brill. Scaling to very large corpora for natural language disambiguation. *Annual Meeting of Association for Computational Linguistics, Toulouse, France, 6-11 July 2001*, pages 26-23.

A. Gegov. *Fuzzy Networks for Complex Systems: A Modular Rule Base Approach*. Springer, Berlin, 2011.

H. Liu, A. Gegov and M. Cocea. *Rule Based Systems for Big Data: A Machine Learning Approach*. Springer, Berlin, 2016.

P. Tan, M. Steinbach and V. Kumar. *Introduction to Data Mining*. Pearson Education, New Jersey, 2006.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |