Noopur Nishikant Zambare
B20ME051

TASK 3

Pros of the paper

In order to analyze the performance a comparison between the performance of RAPIER alone, DISCOTEX without filtering rules on independent data, and DISCOTEX with fully filtered rules has been carried out and discussed in the paper. DISCOTEX using fully filtered rules performs the best, although DISCOTEX without filtering on disjoint data does almost as well. Overall, through these results it could be concluded that data mining's interference has successfully improved the performance of IE.

Working of DiscoTex and how it benefitted the Information Extraction

Documents containing the extracted information, as well as unsupervised data which has been processed by the initial IE system, it is the data that RAPIER would learn to formulate them into supervised set are all used to create a database. Then the work of inner miner is to process this data in order to create the extra set of rules for predictions. These prediction rules are then used during testing to improve the recall of the existing IE system. Firstly the IE works to compose the data into structured one then the RAPIER system comes into play to add the set of rules. But these rules are not sufficient so DiscoTex again generates the new set of rules from the existing data. These rules are little domain specific and improves the prediction at its best possibility.

Cons of the paper

There are certain cons in this text. This might have arrived due to the fact that this paper was published in the year 2000, long 21 years back. With growing years technology has enhanced, algorithms have been amplified with the availability of larger data sets and the training models.

1. **Error with respect to data size**

   According to the modern standards, the size of dataset to analyze the performance is quite low.

   In such cases if outliers are not removed finely then that may give wrong results.

   Small data also decreases the usefulness of the model as it remains much domain

specific if large common attributes are absent. They are highly subjected to random variations and fluctuations.

How the availability of less data could be addressed?

Choose simple models: It has been observed that complex models with many parameters are far more prone to overfitting. In case of Decision Trees, limit the maximum depth. In order to keep the model more conservative, regularization techniques could be used.

2. Problem associated with using decision tress

Decision trees are not fully reliable as just minor variation in the data would result a completely new tree. This problem can be addressed by using an ensemble.

From the above components it could inferred that Data Mining uses ML techniques such as decision trees and not the neural networks. Decision trees models make decisions on the basis of set of rules and in most of the cases these rules are developed by humans. These hand crafted rules might not be fully applicable to the available data since the data is actually gathered from the multiple sources. Although Machine Learning models can learn from data, in the initial stages, they may require some human intervention.

Also decision Tree models could be enhanced by using ensembles.

3. Based on the working algorithm of DiscoTex- C4.5

DiscoTex uses C4.5 rules which uses Decision tress for classification.

C4.5 is an algorithm used to generate a decision tree. The decision trees generated by C4.5 can be used for classification, and for this reason, C4.5 is actually called as a statistical classifier. C4.5 works on the principle of building decision trees from a set of training data.

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized Information Gain .The attribute with the

highest normalized information gain is chosen to make the decision. The C4.5 algorithm then recurses until the base case defined earlier is reached.

Improvement that needs to be done on the basis of algorithm is instead of C4.5, now C5.0 can be used. This would work similar to the previous algorithm but would have few other advantages too.

4. **Use of RAPIER System before applying DiscoTex**
   First of all they used a simpler version of RAPIER system. A full-fledged version might have been more useful for extracting the relevant data. In the paper, they have used a version of RAPIER that employs only the rules which are dependent on words and a little context surrounding a word for forming the set of rules. Whereas the full version of RAPIER system could also use a semantic class information inferred from WordNet which a database of semantic relation of words.

   Rapier's rule representation uses patterns that make use of limited syntactic and semantic information, using freely available, robust knowledge sources such as a part-of-speech tagger and a lexicon with semantic classes.

Critical Analysis of the paper

As the results after the final experiment clearly stated the increase in the prediction accuracy, it could be concluded that there exists a link between data mining algorithms and the simple information extraction that could benefit the final outcome. KDD helped to create relational patterns from the slot fillers that generated more constraints to filter the data with respect to the requirement. Information extraction finds its use in various fields and improving its performance is always helpful to any industry dealing with large data sets.