BENEFITS OF ASSOCIATIVE CLASSIFICATION WITHIN TEXT CATEGORISATION

A. Barnes, T. L. McCluskey and H.Osborne
University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, UK

ABSTRACT

Associative Classification has been successfully employed in many diverse classification problem domains, showing high classification accuracy and adequate computation time relative to the other traditionally used solutions. Despite this, very little research has been conducted with it in the problem area of Text Categorisation and only a small number of approaches presently exist that are based on the concept. This paper aims to highlight the main characteristics of general Text Categorisation problems, provide an overview of the principal drawbacks associated with traditionally employed techniques and outline the benefits of utilising Associative Classification methods as a replacement. The potential disadvantages of the approach are also considered and a range of examples is included for each section in order to present a balanced representation that is unbiased.

Keywords Associative Classification Text Categorisation Document Classification

1 THE TEXT CATEGORISATION PROBLEM

Text categorisation can be described as the assignment of unseen text documents to one or more predefined categories based on their content (Aas et al (1999)) and can be applied to a substantial selection of real world situations.

The problem is very well known and has a long history that dates back to the creation of the first organised text documents, despite this it was only around the late 1980's that research was conducted into text categorisers that had the ability to automatically create classifiers (Sebastiani (1999)). Prior to this the main solution approach was a completely manual process, involving experts with sufficient knowledge in the required subject areas either categorising individual documents or creating generic lists consisting of numerous rules that non-specialists could then follow. Unfortunately these methods are very time consuming and resource intensive, making them extremely expensive and highly impractical, especially when considering the increased use of electronic text databases and mediums that make them readily available, such as the Internet.

Although the generalised problem of Text Categorisation may sound simple, there are a variety of complexities found within the majority of text data that makes it difficult to implement fully automated techniques that give adequate performance. Possibly the most prevalent of these traits is the high dimensional feature space, which arises from the need to tokenise documents into distinct components (break them down into words, word combinations or semantic and syntactic content) allowing them to become indexed and made machine readable. This step, necessary to facilitate automatic learning techniques, typically gives rise to a vast number of features frequently in the order of thousands or tens of thousands even for a relatively small text database (Yang et al (1997)), which can obviously hinder many possible learning approaches due to the computation required.

Other common characteristics include Noise, in the form of irrelevant or redundant features (Gabrilovich et al (2004)), Word Sense Disambiguation (Sanderson (1994)), where a particular word or feature can have multiple dissimilar meanings, and syntactic or semantic content, which relate to the arrangement and intended meaning of text and are of great importance in several text categorisation tasks. Consideration of individual document organisation is also significant as data can be structured, semi-structured and unstructured (Arumugam (2005)), with content composed in a specific controlled style to reflect subject matter, for instance a technical paper, or in a chaotic manner, such as correspondence containing personal thoughts and emotions.

Causing further complication is the issue of text document databases that pose multi-class (Antonie et al (2002)) and, even more notably, multi-label (Thabtah et al (2006)) problems. A multi-class problem is described as one where more than two mutually exclusive categories exist to which individual

documents may be assigned, for example categorising patent manuscripts to their specific author. A multi-label problem is defined as one where each individual document may be assigned to more than a single category at a time, for instance categorising newspaper articles according to the issues and topics they address. Note that multi-labels problems are multi-class problems where the categories are not all considered to be mutually exclusive.

Many real world situations have a need for efficient and high performance text categorisers and though some do not contain multi-class or multi-label elements, such as email classification as 'spam' or 'not spam', the majority do have a requirement for them. Some diverse applications of text categorisation that may not be obviously apparent include (Sebastiani (2005)); patent authorship, news article submissions, speech recognition for telephone routing, caption definition for image classification, survey encoding, question refinement and even automatic essay grading.

Disappointingly, regardless of the vast assortment and diversity of areas requiring text categorisation there is a substantial lack of text document resources available for research purposes, due mainly to privacy issues or generally insufficient or incomplete information. For this reason, most research on text document categorisation is performed on a small selection of benchmark text datasets, which mostly now contain quite dated information and represent only a few of the possible scenarios.

By far the two most prevalent benchmark datasets are the Reuters 21578 corpus (REUTERS) (Lewis et al (2004)) and the OHSUMED corpus (OHSUMED) (Hersh et al (1994)), which both include multi-class and multi-label aspects. Reuters 21578 consists of over twenty thousand newswire stories from 1987 to 1991 and has been extensively refined since its creation with a number of heavily utilised and recognised subsets, such as the "ModApte" Split (Antonie et al (2002)) and the latest incarnation of the entire database labelled "RVC1" (Lewis et al (2004)). By contrast, OSHUMED is a much larger and more complex bibliographical collection of nearly three hundred and fifty thousand references from medical journals, which has had little refinement but still includes a variety of relatively small defined subsets (Yang (1999)).

2 TRADITIONAL TEXT CATEGORISATION APPROACHES

Both the machine learning and information retrieval research circles have heavily investigated the text categorisation problem domain and have consequently developed a large quantity of diverse solution strategies (Aas et al (1999)) (Sebastiani (1999)) (Sebastiani (2005)) (Yang et al (1997)), including neural networks, expert systems, regression models, decision trees, vector models, voting algorithms, rule induction, statistical methods and probabilistic approaches. However most of these are modifications or adaptations of methods found to be effective in other areas of data mining and none are able to handle more than just a small proportion of problem instances and each has weaknesses that inhibit performance.

Despite each solution having its own specific flaws, there are also a number of general disadvantages found throughout the majority of the traditionally used algorithms for text categorisation. Primary among these is the inability to effectively cope with multi-class and, by association, multi-label problems, with the only means of solving them typically being to split the problems and attempt them on a per-class basis and piece the results together afterwards. This virtually always increases the required classifier construction and execution times significantly, as a separate algorithm effectively has to be trained for every class related to a problem and subsequently executed for every document needing classification. Furthermore, a means of combining the output of these independent algorithms must also be employed, which again necessitates additional resources and may have considerable impact on the final results depending on the particular technique chosen and the extent of multi-label characteristics within the data.

In addition to limitations caused by multi-class and multi-label problems, many algorithms also suffer from other defects to varying degrees. As accuracy is widely regarded as the principal performance measure of a text classifier, it is obviously imperative to achieve correct predictions but also equally or perhaps more important to avoid false and potentially misleading ones. Despite some solution approaches being able to attain relatively high classification accuracy in certain situations, there is still much scope for improvement, particularly in instances of multi-label problems where it is crucial that documents are not under or over allocated categories. Presently it proves very challenging to avoid this incorrect consignment of category density due to the way most traditional solutions divide multi-

label problems into sub tasks, as there are few options available when piecing the individually solved parts back together. One way to bypass this dilemma is to simply ignore it, but that could produce an unfavourable imprecision in perceived performance and create a bias towards classifiers that allocate excessive categories, which in turn may lead to them being erroneously selected for tasks they are not really suited to. Fortunately, directly in acknowledgement of the problem, performance is often assessed by means that take both "precision" and "recall" values (Sebastiani (1999)) into account, which track if documents are correctly classified and whether categories have been over assigned.

In contrast to accuracy, the time taken to construct and execute a classifier is persistently neglected as it is generally thought that hardware advancements make its consideration unnecessary, resulting in an extensively accepted notion of 'effectiveness over efficiency' (Sebastiani (1999)). However, though some methods might exhibit reasonable efficiency they are also likely to have exponential computation time in relation to both problem size and complexity, making them decidedly impractical for most real world applications. Even those that do demonstrate adequate efficiency still tend to suffer from another drawback of many traditional approaches, which is the tendency to construct classifiers that function as a 'black box', meaning it is not easily interpreted or understandable by humans. For situations that require entirely automated systems with no human interaction this would obviously not be an issue, but a sizable number of applications benefit greatly from classifiers that have easily decoded outputs that can be augmented or modified with minimal effort.

Two of the consistently best performing categorisation techniques, with regard to classification accuracy, are K-Nearest Neighbour (KNN) (Yang et al (1999)) (Aas et al (1999)) and Support Vector Machines (Yang et al (1999)) (Joachims (1998)). Although both of these are frequently superior to other procedures by a notable margin and contain several potential advantages, they still contain inherent weaknesses that render them unsuitable in certain circumstances.

KNN is a comparatively fast approach that uses statistical information extracted from the data and requires little or no offline training in the form of a classifier construction phase, due to parameter adjustments that can be automatically fine tuned and a lazy learning approach (Yang et al (1999)). During classification each test document is checked against 'k' similar documents from the training set, which are determined according to some chosen similarity measure, and the categories assigned to those documents are ranked following set criteria to form a prioritised list of potential categories. The foremost limitations of KNN are the choice of parameter thresholds and the similarity measure, which both affect the time taken to discover similar documents, classification accuracy and interpretability of the output.

Using a slightly different process, the fully automated SVM requires no user interaction and employs an integrated kernel (Joachims (1998)) to create vectors from the features of unclassified documents, which are then compared to prototype vectors previously derived from training data. These prototype vectors each relate to a single category and are formed by finding the maximal margin hyper-plane that separates all features from training documents assigned to a category from all of those of documents not assigned to it. Major restrictions with this approach include the memory requirements, which expand exponentially with data growth, its inefficient handling of multi-class problems, as separate test and prototype vectors are needed for each class, and user comprehension of all the resultant vectors. Selection of the kernel to incorporate within the SVM is commonly also a significant factor that depends heavily on the type and characteristics of the problem being solved, however it has been shown that ones based on simple linear equations work well in this setting.

3 ASSOCIATIVE CLASSIFICATION APPROACHES

Associative Classification (Liu et al (2000)) (Thabtah (2005)) is a union between the data mining tasks of association rule mining (Janssens et al (2003)) and classification rule mining (Freitas (2000)) and was first introduced as a phrase in 1998 (Liu et al (1998)). The main concept is the application of association rule mining techniques on classification databases as opposed to transactional databases, which do not contain class labels, by focusing exclusively on a distinct subset of rules named "Class Association Rules" (CAR). These are able to portray relationships between the features of a database item and its correlating class labels by extracting specialist rules that are formed with only data attributes as the antecedent and only class labels as the consequent (Thabtah (2007)).

Due to the inherent exhaustive nature of association rule mining techniques, classification algorithms utilising them typically extract more rules than traditional methods, which often leads to the discovery of some that are otherwise concealed and unobtainable. By the use of these supplementary hidden rules it has been demonstrated that it is frequently possible to achieve highly competitive or in some cases significantly increased classifier accuracy (Thabtah (2007)). This combination of enhanced rule discovery and the ability to interchange the association rule mining element of any algorithm for one that is more efficient, without the risk of degrading accuracy (Freitas (2000)), makes it an extremely promising classification solution. Adding further to this appeal is the high level of output interpretability, which follows a readily understandable convention and can be easily adapted by humans or input into other systems.

Another prominent advantage offered is the capacity to handle multiple data features simultaneously for the fabrication of rules, overcoming the shortfalls related to dealing with only single variables or classes at once, something that seriously affects the composition of true multi-class and multi-label classifiers (Thabtah et al (2006)). This consequently allows a global perspective of problems containing these characteristics to be achieved, as they are no longer broken into small isolated sub divisions, which permits much greater flexibility when considering unstructured or semi-structured data. Not only does this potentially yield superior rules, each able to account for multiple attributes and classes, but it also means fewer scans of the database are required as the algorithm needs just a single iteration regardless of the number of categories being dealt with.

Surprisingly, though it seems well suited to the task, little work has taken place using associative classification within the text categorisation domain and there are only a limited number of solutions based around the concept. Even then, these approaches have only trivial exposure to the research community, typically lacking detailed results and sufficient information to demonstrate their full potential or permit duplication for extended assessment.

Association Rule based Classifier All Categories (ARC-AC) (Antonie et al (2002)), one of the earliest document classification methods to contain an association rule mining element, is based on the Apriori algorithm (Thabtah (2005)) and uses basic rule ranking and database coverage for pruning. Derived directly from this is Association Rule based Classifier By Categories (ARC-BC) (Antonie et al (2002)), which has the same composition but instead of adopting a global approach it extracts rules by treating each category separately and combining them afterwards, similar to traditional methods. Both algorithms attain accuracies comparable to the best traditional solutions and have fast classifier construction and execution times, whilst providing easily interpreted outputs.

Association Rule based Text Classifier (ARTC) (Buddeewong et al (2005)), an enhancement of ARC-BC, was developed in order to overcome some of the issues encountered during the processing of features that overlap multiple categories. It does this by creating two independent frequent item-sets, one exclusively containing features with no overlap and one containing solely overlapping features, which are combined into candidate frequent item-sets by using two separate methods. Although it is reported to have a notably greater accuracy than unmodified ARC-BC, it is difficult to validate this as inadequate experimental results are available and much of the information relating to their generation is absent.

Moving away from slight modification of existing algorithms, the Negated Words (NeW) classifier (Baralis et al (2006)) creates more specialised rules by using the chi square test (Yang et al (1997)) to prune low quality rules and by considering the absence of words as part of the rule antecedent. This combination of positive and negative examples as part of the rule helps to avoid incorrect document classification and rarely has a detrimental effect, which consequently increases overall accuracy with marginal overheads.

Despite the many benefits to be gained through the use of associative classification, it does also have a number of negative aspects that need to be addressed (Li et al (2001)). Perhaps the greatest of these is caused directly by the exhaustive rule detection, which can generate excessively large rule sets that expand exponentially as the problem feature space grows. Unless this is suitably managed with appropriately aggressive pruning, it may inhibit the effectiveness and scalability of an algorithm and cause issues when attempting to select the optimum classification rules. Aside from this, the discovery process is also sensitive to the user controlled parameters of support and confidence, both of which can give a desired accuracy to speed ratio if tuned properly or have drastic effects if not. A

poorly chosen value for the support threshold would mean the production of either too few or too many rules, while an incorrect confidence threshold would influence the amount of under and over-fitting that occurs within the training data.

4 CONCLUSION/SUMMARY

The problem of Text Categorisation has a long history and applies to a vast array of real world applications, yet it is only since the late 1980's that research has been conducted to move from the resource intensive and time consuming manual process to an automated one. However, due to the habitual characteristics of text document datasets this is not an easy step, as they consistently contain several complexities such as high problem dimensionality, inherent noise and multi-class or multi-label aspects.

In an attempt to overcome these obstacles, a multitude of varying traditional algorithms have been adopted from other data mining areas and applied to this domain, but though they each display some advantages most also share a persistent set of limiting weaknesses. Primary amongst these is the inability to effectively cope with non-binary problems that contain more than two distinct class labels, especially when the labels are not mutually exclusive to each other, which are both common traits of real world applications.

Approaches utilising Associative Classification are able to overcome this inadequacy, along with many of the other drawbacks regularly encountered, by providing a true multi-class and multi-label solution whilst still providing interpretable results and retaining the flexibility to deal with other problem issues. Even so, regardless of these potential benefits, a number of shortcomings inherent with the method need consideration, including the substantial quantity of rules created by the exhaustive search technique and the proper tuning of the user defined support and confidence parameters.

5 FUTURE WORK

The current main focus is the completion of a prototype framework specialised for the processing of text document data and the creation and execution of varied text categorisation solutions, including a mixture of traditional approaches and those based around the Associative Classification concept. A series of informational papers are also being developed inline with the framework, taking the format of surveys and white papers that describe the stages involved within the framework and allow its use, modification or reconstruction by other researchers that may find it beneficial.

Upon completion of the prototype an assortment of different approaches, including the best performing traditional and associative classification techniques applied to the problem area, will be reproduced and used as a baseline set of results for comparison of novel associative classification algorithms. These algorithms are primarily aimed at improving upon the existing classification accuracy presently obtainable for a number of diverse text document databases and then, time permitting, will be further enhanced to provide superior computational times for both classifier construction and execution.

REFERENCES

AAS K. and EIKVIL L. (1999), *Text Categorisation: A Survey*, Technical Report, Norwegian Computing Center

ANTONIE M. and ZAANE O. (2002), *Text Document Categorisation by Term Association*, IEEE International Conference on Data Mining, pp. 19-26

ARUMUGAM S. (2005), *Classification Techniques for Categorization of Hypertext Documents*

BARALIS E. and GARZA P. (2006), *Associative Text Categorization Exploiting Negated Words*, Symposium on Applied Computing, pp. 530-535

BUDDEEWONG S., KREESURADEJ W. (2005), *A New Association Rule-Based Text Classifier Algorithm*, ICTAI, 17th IEEE International Conference on Tools with Artificial Intelligence, pp. 684-685

FREITAS A. A. (2000), *Understanding the Crucial Differences between Classification and Discovery of Association Rules – A Position Paper*, SIGKDD Explorations, Vol. 2, pp. 65-69

GABRILOVICH E. and MARKOVITCH S. (2004), *Text Categorisation with Many Redundant Features: Using Aggressive Feature Selection to Make SVMs Competitive with C4.5*, ICML

HERSH W., BUCKLEY C., LEONE T. J. and HICKMAN D. (1994), *OHSUMED: An Interactive Retrieval Evaluation and New Large Test Collection for Research*, 17th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, Pp. 192-201

JANSSENS D., WETS G., BRIJS T. and VANHOOF K. (2003), *Integrating Classification and Association Rules by Proposing Adaptations to the CBA Algorithm*

JOACHIMS T. (1998), *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*, 10th European Conference on Machine Learning, pp. 137-142

LEWIS D. D., YANG Y., ROSE T. G. and LI F. (2004), *RVC1: A New Benchmark Collection for Text Categorization Research*, Journal of Machine Learning Research, Vol. 5, pp. 361-397

LI W., HAN J. and PEI J. (2001), *CMAR: Accurate and Efficient Classification Based on Multiple Class-Association Rules*, ICDM, pp. 369-376

LIU B., HSU W. and MA Y. (1998), *Integrating Classification and Association Rule Mining*, Knowledge Discovery and Data Mining, pp. 80-86

LIU B., MA Y. and WONG C. K. (2000), *Improving an Association Rule Based Classifier*, Principles of Data Mining and Knowledge Discovery, pp. 504-509

OHSUMED Database – http://trec.nist.gov/data/t9_filtering/README

REUTERS Database – http://kdd.ics.uci.edu//databases/reuters21578/README.txt

SEBASTIANI F. (1999), *A Tutorial on Automated Text Categorisation*, ASAI-99, 1st Argentinean Symposium on Artificial Intelligence, pp. 7-35

SEBASTIANI F. (2005), *Text Categorization*, Text Mining and its Applications to Intelligence, CRM and Knowledge Management, pp. 109-129

THABTAH F. (2005), *Multiple Labels Associative Classification Mining*, PhD Thesis

THABTAH F., COWLING P., PENG Y. (2006), *Multiple Labels Associative Classification*, Knowledge and Information Systems, Vol. 9, Issue 1, pp. 19-129

THABTAH F. (2007), *A Review on Associative Classification Mining*, Journal of Knowledge Engineering Review, Vol. 22, No. 1, pp. 37-65

YANG Y. and PEDERSEN J. O. (1997), *A Comparative Study on Feature Selection in Text Categorization*, 14th International Conference on Machine Learning, pp. 412-420

YANG Y. (1999), *An Evaluation of Statistical Approaches to Text Categorization*, Journal of Information Retrieval, Vol. 1, pp. 60-90

YANG Y. and LIU X. (1999), *A Re-Examination of Text Categorization Methods*, 22nd Annual International SIGIR, pp. 42-49

SANDERSON M. (1994) (re-published 1997), *Word Sense Disambiguation and Information Retrieval*, 17th annual international ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 49-57