Genetic Algorithm based Feature Selection for Ontology based Information Retrieval of Semi Structure Data

N. Vanjulavalli[1], Dr. A. Kovalan [2]

[1.] Research Scholar,Department of Computer Science and Applications,PMU, Vallam, Thanjavur
[2.] Assistant Professor (S.S), Department of Computer Science and Applications,PMU, Vallam, Thanjavur.
vanjulavallisn@gmail.com

Abstract: The increasing volume of web pages in World Wide Web in the form of free-text makes information retrieval difficult. The retrieval is more challenging due to the ambiguous nature of the unstructured information found in these pages. Ontologies help to overcome the disambiguate nature of the natural language by the use of standard terms that relate to specific concepts. Thus, the knowledge of ontology is used to match object and queries based on semantics improving information retrieval. In this paper, the features from the web pages are extracted based on ontology and semantics of the XML tags. Genetic Algorithm is applied for selecting optimal subset of features based on correlation. Experimental results for the proposed feature extraction method demonstrate the effectiveness of the optimization of the feature selection.
[N. Vanjulavalli, A. Kovalan. **Genetic Algorithm based Feature Selection for Ontology based Information Retrieval of Semi Structure Data.** *Life Sci J* 2013;10(7s):516-521]. (ISSN: 1097-8135). http://www.lifesciencesite.com. 80

Keywords: Information retrieval (IR), World Wide Web, Ontology, Feature Selection, Genetic Algorithm, Bagging

1. Introduction

Ontology refers to the structured representation of the domain knowledge which includes defining of classes, relations and functions among the objects [1, 2]. Ontology models the relationship between the concepts and objects for a domain. Information retrieval for semi structured data such as web pages is challenging due to the ambiguous nature of the unstructured information found in these pages. During information retrieval, words in natural language may have different meanings depending on the context leading to inefficient retrieval [3]. In ontology, the context of vocabulary is represented and constrained in the ontology model, thus, overcoming the disambiguate meanings of words in the free text.

Ontology is a formal and explicit specification of a shared conceptualization [4, 5] and is specific for particular domains. Modelling the domain's concepts through ontology helps the process of information retrieval [6]. For example, educational institute ontology defines concepts like courses, faculty and students which are useful for information extraction. This forms the basis of ontology-based information extraction. Ontology-Based Information Extraction (OBIE) is a fast evolving information extraction sub field. Here, ontologies are used by information extraction process with the output being similarly represented through ontology. Ontologies are essential in semantic web applications as it provides shared knowledge about the real world objects leading to reusability and interoperability among different modules. For that reason, the ontology quality is critical in any semantic application. The knowledge of ontology is used to match object and queries based on semantics improving information retrieval.

Extensible Markup Language (XML) is widely used to represent data on the internet. Advanced query engines allow users to exploit data in XML documents. New semi structured data models and query languages were proposed for this purpose [7, 8]. XML data is self-describing, programs can interpret data meaning. The data can be filtered based on content, restructure it to suit applications etc. The application of ontologies face a number of challenges when applying machine learning techniques on the features extracted. The features could be extracted either from concepts of ontologies or Natural Language Processing. This paper explores the use of ontology to match object and queries semantically.

Features extracted are the key for achieving good classification. Feature selection is widely used in data mining methods to reduce the number of features, remove redundant data which leads to improved performance of the classification algorithm. A subset is selected from the original features during feature selection based on some evaluation criteria to obtain the optimal feature subset. As the number of features increase, finding the optimal feature subset is NP-hard. Existing methods for feature selection such as a filter or wrapper methods do not essentially produce the optimal subset. Genetic Algorithm is widely used as optimizing tool, in this paper; it is used to search for the optimal subset of features.

The features are extracted from XML documents based ontologies concepts and Natural Language Processing. Genetic algorithm is used to find the optimal feature subset. The 4 Universities Dataset includes internet pages from computer science departments of leading universities which evaluate the proposed method. The following sections detail some of the related works available in the literature, proposed methodology and the experimental results.

2. Related Works

For the purpose of indexing and retrieving documents conventional information retrieval systems depends on the textual keywords. In queries, while using diverse keywords to describe the same concept, the keyword-based retrieval might provide incorrect and imperfect results. Instead of being syntactic, the related keywords relationship might be semantic that necessitates access to complete human world knowledge in order to capture it. Implementing thesauri that is developed manually, concept-based retrieval approaches have tried to address these complications by depending on the data with term co-occurrence, or by obtaining concepts from a corpus and the relationships between latent words. Egozi et al., [9] proposed a novel concept-based retrieval method on the basis of Explicit Semantic Analysis (ESA), an approach to augment keyword-based text representation using concept-based characteristics, obtained from large human knowledge repositories like Wikipedia in an automatic way. Text features are extracted automatically by the proposed method, and it was revealed that complication was observed for high-quality feature selection in this setting as the main emphasis was on retrieval. But, the conventional feature selection approaches cannot be implemented due to the lack of labeled data. Therefore, using self-generated labeled training data novel approaches are proposed. On evaluating the resultant system on different TREC datasets illustrates better performance over the other existing state-of-the-art results.

Both in academia and industry, Ontology based Retrieval System is widely developed and studied. But, the tribulations experienced by most of the systems are: The robust hierarchical classifiers training are necessitated to construct the mappings among documents and concepts in ontology and the documents distribution is ignored by the classical Browsing Hierarchical System. Thus, for users browsing documents such concepts becomes unpractical and consumes more time. Hence, organizing documents in the browsing system becomes more complicated and further splitting of these concepts into sub-categories is made compulsory. In order to develop a realistic and precise Hierarchical Browsing System, Nanhong Ye et al., [10] proposed an adaptive Hierarchical Browsing System framework that constructs an adaptive Ontology based Hierarchical Browsing System for *CiteSeer*^{x}. In this architecture, initially, the documents are classified into present predefined concepts of ontology and using different datasets of *CiteSeer*^{x} their performance is compared by examining the supervised learning methods. Next, to add novel clusters to the present browsing hierarchy, an empirical estimation unsupervised learning approaches is performed. The efficiency and efficacy of the proposed approach is revealed by the experimental results on *CiteSeer*^{x} corpus.

Tuominen et al., [11] used the ontologies published in the ONKI Ontology Service and presented an ontology-based query expansion Widget. To improve the functioning, the widget can be incorporated into a web page, for instance: in the search system of a museum catalogue to give query expansion functionality in order to improve the page. Using general, spatio-temporal and domain specific ontologies, the proposed system was experimented. Few challenges were faced while using the ONKI widgets with the illustration search interface for the Kantapuu.fi system. The expanded query string becomes very long and inconvenient when a query concept consists of many sub concepts and also when an additional concept URIs/labels of the sub concepts is added to the query. As the database system, HTTP server or any other software components used restricts the length of the query string leads to cause some tribulations. The system may not perform accurately or the response times are increased due to the length of the queries.

Koopman et al., [12] suggested a new method developed on the basis of concept matching instead of keyword matching for the purpose of searching in electronic medical records. The SNOMED-CT ontology defines that the documents and queries are modified from their term-based originals into medical concepts. A real-world set of medical records is used for estimation. The results show that the keyword baseline is outperformed by the proposed concept-based approach in a Mean Average Precision of 30%. Additionally, for future improvement in inference based search systems, the concept-based approach gives an exact framework to deal the medical data.

3. Methodology

The 4 Universities Dataset

The 4 Universities Dataset includes WWW-pages from computer science departments of leading universities: Cornell, Texas, Washington, Wisconsin and 4,120 miscellaneous pages from other universities. It was collected in January 1997 by the

CMU text learning group's World Wide Knowledge Base (Web->Kb) project [13]. The dataset consist of 8,282 pages and were manually classified into seven categories (student, faculty, staff, department, course, project and other). The files are organized into a directory structure, one directory for every class. Each of the directories includes 5 subdirectories, one for each of the 4 universities and one for miscellaneous pages. The directories in turn contain Web-pages.

Feature Extraction

Features are extracted from the web pages using stemming, stop words, and Inverse Document Frequency (IDF). IDF measures a word's ability to discriminate between documents [14]. Inverse Document Frequency (IDF) is measures a word's importance and is defined as the logarithm of the ratio of documents in a collection to the number of documents containing given words, thus rare words possess high IDF value and common words low value IDF. Document and query are represented as vectors in a high dimensional space. Similarity measures between keywords and the document is computed and ranking is based on them.

The stop word list is a list of non-significant words to be removed from a document before indexing. Stop word list is for words serving no retrieval purpose but used frequently to compose documents. Non-significant words cause inefficient retrieval as they fail to discriminate between relevant and non-relevant documents. Stop word list is generally made up of many pronouns, articles, prepositions and conjunctions. Words like the, a, of, for, with etc., are stop words.

Stemming enhances retrieval effectiveness by removal of inflectional and derivational suffixes to conflate word variants into the same stem/ root. It is assumed that words with similar stem refer to the same concept and therefore can be indexed under the same form. Stemming removes inflectional suffixes or, for English, this conflates singular/plural word forms and removes past participle ending «-ed» and gerund or present participle ending «-ing».

Let frequency be denoted by $freq(x,a)$, as it expresses the number of occurrences of the term a in a document x. The term-frequency matrix $TF(x,a)$ measures term a association regarding given document x. $TF(x,a)$ is assigned zero when document does not contain the term, and a number otherwise. The number can be set as $TF(x,a) = 1$ when term a occurs in document x or uses relative term frequency which the frequency versus total occurrences of all document terms. Another measure is **inverse document frequency (IDF)**, representing a scaling factor. If term a occurs frequently in documents, its importance is scaled down due to lowered discriminative power. The $IDF(a)$ is defined as follows:

$$IDF(a) = \log \frac{1+|x|}{x_a}$$

x_a is the set of documents containing term a. Similar documents have same relative term frequencies. Similarity is measured among a document set/between a document and query. Cosine measure locates document similarity [15]; cosine measure is got by:

$$sim(v_1, v_2) = \frac{v_1 . v_2}{|v_1| \ |v_2|}$$

where v_1 and v_2 are two document vectors, $v_1 . v_2$ defined as $\sum_{i=1}^{a} v_{1i} v_{2i}$ and $|v_1| = \sqrt{v_1 . v_1}$.

In the proposed features extraction, the features are extracted based on the ontology and feature selection is achieved by genetic algorithm. A concept based tree structure is built on a generalisation/specialisation relationship to conceptualization the domain. Browsing knowledge is made easier if the conceptual architecture of the knowledge based is identified as a whole and information is accessible by intra conceptual hierarchical links during browsing. Thus, when browsing in a vast information base, data mapping provides interesting solutions in representing the data [16]. This is also applicable to semantically annotated knowledge bases resulting in concepts tree structure. The concepts are organized into a taxonomy tree where each node represents a concept and every concept a specialization of its parent. Figure 1 shows simplified taxonomy tree for Computer Science (CS) department domain.

On extraction of features based on ontology and semantics, feature selection is applied to reduce the number of features used for classification of the web pages. Genetic Algorithm (GA), an optimization technique, is applied to find the optimal subset of features. GAs are based on the genetic and natural selection principles and are efficient adaptive search techniques [17]. GA is initialized with the creation of the population of individuals, where an individual represents a sample of space to be searched. For selection of the optimal subset, the individuals in the population are encoded as a subset of the features. Each individual is evaluated on the basis of overall fitness.

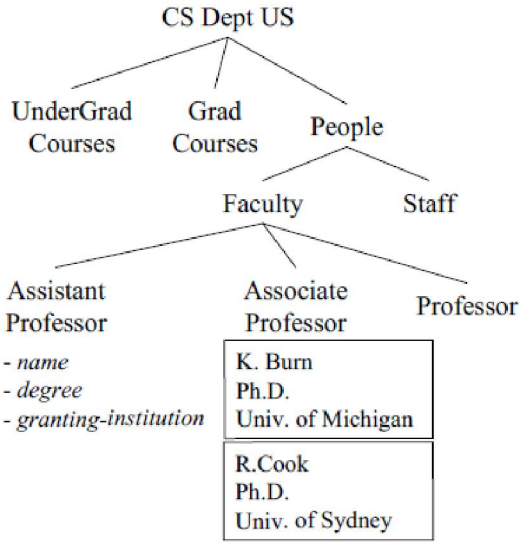

Figure 1: Sample Ontology Tree

In this study, the correlation between the features is used as the measure of the fitness.

$$cr = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\left(\sum X^2\right) - \left(\sum X\right)^2}\sqrt{N\left(\sum Y^2\right) - \left(\sum Y\right)^2}}$$

Lower the correlation, higher is the fitness. Thus, individual with lower correlation is more fit. New individuals (or subsets) are created by selecting two individuals with high fitness. The child produced retains many of the features of the parents leading to a population of improved fitness.

New individuals for the next generation are created by using genetic operators (crossover and mutation). Crossover forms children by crossing over with selected parents, forming similar individuals. Mutation helps to inject new information into the population by randomly changing one or more components in an individual. Thus, the GA starting from an initially unknown search space, through each iteration moves to promising subspace [18]. The fitness is evaluated for each generation, and the cycle continues until the termination criteria is satisfied. Figure 2 shows the process of the selection of the optimal subset of features using GA.

Bagging

The Bagging algorithm (Bootstrap aggregating) votes classifiers generated by different bootstrap samples [19]. A Bootstrap sample make certain that uniform generation of sampling m instances from the training set with replacement is achieved. T bootstrap samples $B_1,\ldots B_T$ are produced and classifier C_i is built from each bootstrap sample B_i. A final classifier C* is built from $C_1,\ldots C_T$ whose output is most predicted class by sub-classifiers, with arbitrarily broken ties.

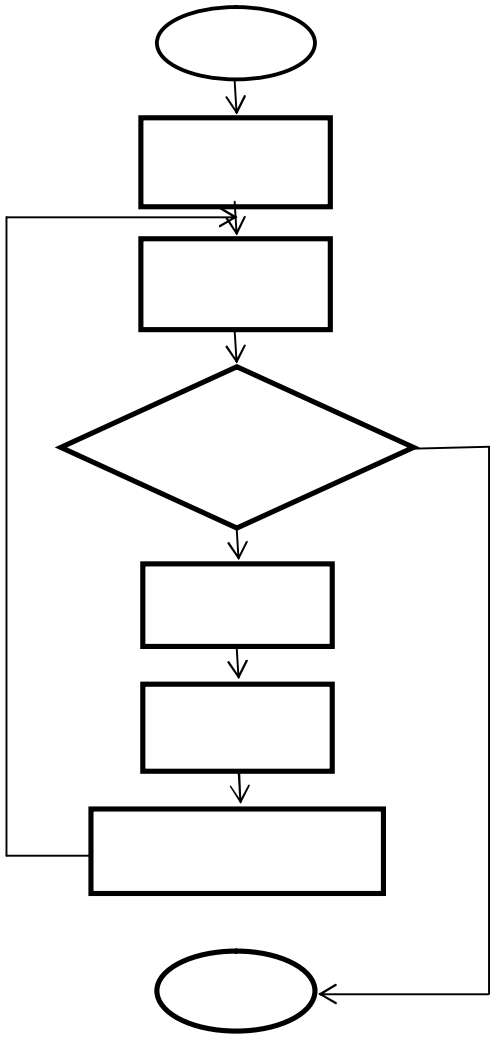

Figure 2: Proposed Genetic based Feature Selection

Bagging algorithm is as follows [20]:
Inputs: training set S, Inducer I, Number of bootstrap samples T
for i= 1 to T {
S' = bootstrap sample from (sample with replacement)
Ci= I(S')
}

$$C^*(x) = \arg\max_{y \in Y} \sum_{i:C_i(x)=y} 1$$

Output: classifier C*

In this study, the Bagging is done with REPtree, BFtree, J48, and CART.

4. Results and Discussion

The proposed genetic based feature extraction for web page classification is assessed

using the 4 Universities Dataset and compared with IDF feature extraction method. Classification accuracy, Recall and precision are measured for both proposed and IDF techniques. The accuracy, precision, recall and f measure are computed as follows:

Accuracy (%) = (TN + TP) / (TN + FN + FP + TP)

$$precision = \frac{TP}{TP + FN}$$

$$recall = \frac{TP}{TP + FP}$$

$$f\ Measure = \frac{2 * recall * precision}{recall + precision}$$

where TN (True Negative) = Number of correct predictions that an instance is invalid

FP (False Positive) = Number of incorrect predictions that an instance is valid

FN (False Negative) = Number of incorrect predictions that an instance is invalid

TP (True Positive) = Number of correct predictions that an instance is valid

Bagging with various decision trees (REPtree, BFtree, J48, and CART) are the classifiers used. The experimental results obtained are detail in the following tables and figures. Table 1 and Figure 2 details the classification accuracy and root mean squared error obtained for IDF and proposed genetic feature extraction.

Table 1: Classification Accuracy and Root Mean Squared Error

Method Used	Classification Accuracy %	RMSE
Bagging-REPtree-IDF	0.71	0.32
Bagging-BFtree-IDF	0.74	0.31
Bagging-J48-IDF	0.73	0.31
Bagging-CART-IDF	0.77	0.31
Bagging-RREPtree - Proposed genetic feature extraction	0.87	0.23
Bagging-BFtree-Proposed feature extraction	0.88	0.21
Bagging-J48-Proposed feature extraction	0.92	0.19
Bagging-CART-Proposed feature extraction	0.85	0.22

It is observed from Figure 3, that the proposed genetic feature extraction achieves better classification accuracy than the traditional IDF features. The Root Mean Squared Error is lower for the proposed method. The precision, recall and f measure for the different methods is shown in Table 2 and Figure 4 and 5 shows the precision, recall and f measure respectively.

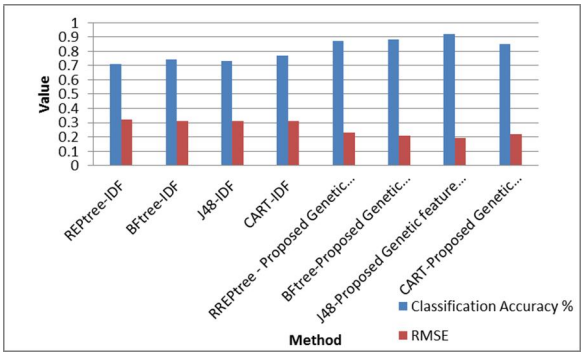

Figure 3: Classification Accuracy and Root Mean Squared Error

Table 2: Precision, Recall and F Measure

Method Used	Precision	Recall	f Measure
Bagging-REPtree-IDF	0.715	0.71	0.709
Bagging-BFtree-IDF	0.747	0.74	0.741
Bagging-J48-IDF	0.725	0.73	0.725
Bagging-CART-IDF	0.777	0.77	0.772
Bagging-RREPtree-Proposed feature extraction	0.882	0.87	0.873
Bagging-BFtree-Proposed feature extraction	0.894	0.88	0.883
Bagging-J48-Proposed feature extraction	0.926	0.92	0.92
Bagging-CART-Proposed feature extraction	0.876	0.85	0.856

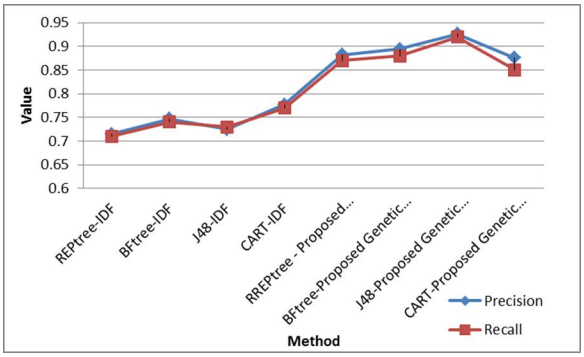

Figure 4: Precision and Recall

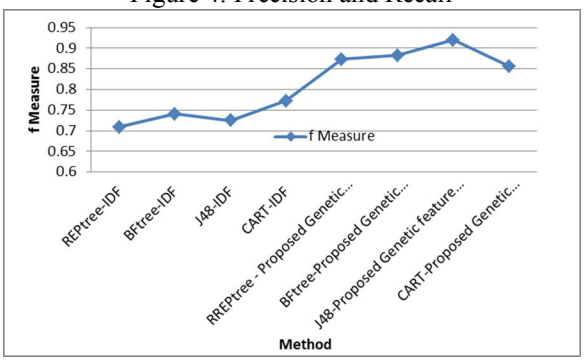

Figure 5: F Measure

The F Measure score combines recall and precision ensuring that an F Measure score has values within the interval [0, 1]. The F Measure is less when number of relevant documents retrieved is less and high score is achieved when retrieved documents are relevant. F Measure has a high value only when precision and recall are also high. The maximum value for F Measure gives the best possible compromise between recall and precision. It is observed from Figure 5 that the combination of Bagging with J48 and the proposed genetic feature extraction achieves the highest F Measure score.

5. Conclusion

Ontology-Based Information Extraction is a widely researched information extraction sub field. In this paper, ontologies are used for information extraction process. Features are extracted using information retrieval approaches such as IDF and proposed ontology based features. The extracted features are processed using genetic algorithm to find optimal feature subset which is used as the input for the classifiers. The feature subset is classified using Bagging with various decision trees (REPtree, BFtree, J48, and CART). The experimental results show that proposed feature extraction improves the precision and recall satisfactorily.

References

1. Cheng, C. K., Pan, X., & Kurfess, F. (2004). Ontology-based semantic classification of unstructured documents. In Adaptive Multimedia Retrieval (pp. 120-131). Springer Berlin Heidelberg.
2. Vallet, D., Fernández, M., & Castells, P. (2005). An ontology-based information retrieval model. In The Semantic Web: Research and Applications (pp. 455-470). Springer Berlin Heidelberg.
3. Pan, X., & Assal, H. (2003, October). Providing context for free text interpretation. In Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003 International Conference on (pp. 704-709). IEEE
4. T. R. Gruber, A translation approach to portable ontology specifications, Knowledge Acquisition 5(2) (1993) 199-220.
5. R. Studer, V.R. Benjamins and D. Fensel, Knowledge Engineering: Principles and methods, Data Knowledge Engineering 25(1) (1998) 161-197.
6. Wimalasuriya, D. C., & Dou, D. (2010). Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 36(3), 306-323.
7. He, J. S. K. T. G., & Naughton, C. Z. D. D. J. (2008). Relational databases for querying XML documents: Limitations and opportunities. 20.453J / 2.771J / HST.958J Biomedical Information Technology Fall 2008.
8. Melton, J., & Buxton, S. (2011). Querying XML: XQuery, XPath, and SQL/XML in context. Morgan Kaufmann.
9. Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. ACM Transactions on Information Systems (TOIS), 29(2), 8.
10. Ye, N., Gauch, S., Wang, Q., & Luong, H. (2010, October). An Adaptive Ontology Based Hierarchical Browsing System for CiteSeerx. In Knowledge and Systems Engineering (KSE), 2010 Second International Conference on (pp. 203-208). IEEE.
11. Jouni Tuominen, Tomi Kauppinen, Kim Viljanen, and Eero Hyv¨onen. Ontology-based query expansion widget for information retrieval. In Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), 6th European Semantic Web Conference (ESWC 2009), May 31 - June 4 2009.
12. Koopman, B., Bruza, P., Sitbon, L., & Lawley, M. (2012). Towards semantic search and inference in electronic medical records: an approach using concept-based information retrieval. Australasian Medical Journal.
13. Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine learning, 39(2), 103-134.
14. Salton, G., & McGill, M. J. (1986). Introduction to modern information retrieval.
15. Steinbach, M., Karypis, G., & Kumar, V. (2000, August). A comparison of document clustering techniques. In KDD workshop on text mining (Vol. 400, pp. 525-526).
16. Doan, A., Madhavan, J., Domingos, P., & Halevy, A. (2004). Ontology matching: A machine learning approach. Handbook on Ontologies in Information Systems, 397-416.
17. Holland, J. H.,. "Adaptation in Natural and Artificial Systems," University of Michigan Press, Ann Arbor, MI., 1975.
18. De Jong, K., "Learning with Genetic Algorithms : An overview," Machine Learning Vol. 3, Kluwer Academic publishers, 1988.
19. Breiman, L. (1996b). Bias, variance, and arcing classifiers. Tech. rep. 460, Department of Statistics, University of California, Berkeley, CA.
20. Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36 (1/2), 105{139.

1/8/2013