ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 3, May 2013

An Efficient Information Retrieval from Domain Expert Using Active Learning with Generalized Queries

S .Brintha Rajakumari[*1]

Department of CSE, Bharath Institute of Science and Technology, Chennai, Tamil Nadu, India[*1]

ABSTRACT: *In* recent years, domain-driven data mining (D3M) has received extensive attention in data mining. Unlike the traditional data-driven data mining, D3M tends to discover actionable knowledge by tightly integrating the data mining methods with the domain-specific business processes. However, in most cases, the domain specific actionable knowledge cannot be discovered without the support of domain knowledge, mainly provided by human experts. Thus, the human-machine-cooperated interactive knowledge discovery process is widely applied in real-world applications. Active learning can integrate the automated learning algorithm with the domain experts. The main aim of the paper is to get the information from domain experts to the generalized queries with don't care attribute using data mining with addition off Active Learning with Generalized Queries Algorithm.

KEYWORDS: Active Learning, Domain Driven Data Mining, Information Retrieval.

I. INTRODUCTION

With the assistance of a domain expert, active learning can often select or construct fewer examples to request their label to build an accurate classifier. However, previous works of active learning can only generate and ask specific queries. In real-world applications, the domain experts or oracles are often more readily to answer "generalized queries" with don't-care attributes. The power of such generalized queries is that one generalized query is often equivalent to many specific ones. However, overly general queries are not good as answers from the domain experts or oracles can be highly uncertain, and this makes learning difficult.

in this paper, a novel active learning algorithm is used that asks good generalized queries. it then, extends an algorithm to construct new, hierarchical features for both nominal and numeric attributes. it demonstrate experimentally that the new method asks significantly fewer queries compared with the previous works of active learning, even when the initial labeled data set is very small, and the oracle is inaccurate[1]. this method can be readily deployed in real-world data mining tasks where obtaining labeled examples is costly. the main aim of the paper is to get the information from domain experts to the generalized queries with don't care attribute using data mining with addition off agq algorithmformat & style

A .Domain Driven Data Mining

Data mining and knowledge discovery [2] (data mining or KDD for short) has emerged to be one of the most vivacious areas in information technology in the last decade. It has boosted a major academic and industrial campaign crossing many traditional areas such as machine learning, database, and statistics, as well as emergent disciplines, for example, bioinformatics. As a result, KDD has published thousands of algorithms and methods.

Data mining is a powerful paradigm of extracting information from data. It can help enterprises focus on important information in their data warehouse. Data mining is also known as Knowledge Discovery in Databases (KDD). It involves the extraction of hidden pattern to predict future trends and behaviors which allow businesses to make proactive, knowledge-driven decisions.

Domain driven data mining [3] involves the study of effective and efficient methodologies, techniques, tools, and applications which can discover and deliver actionable knowledge that can be passed on to business people for direct decision-making and action-taking.

A key concept in D3M that is highlighted is Actionable Knowledge Discovery (AKD). It involves and synthesizes domain intelligence, human intelligence and cooperation, network intelligence and in-depth data intelligence to define, measure, and evaluate business interestingness and knowledge action ability. The authors stressed the importance of AKD as an important concept for bridging the gap between technical-based approaches and business impact-oriented expectations on patterns discovered from data mining.

B .Active Learning

The primary goal of machine learning is to derive general patterns from a limited amount of data. The majority of machine learning scenarios generally fall into one of two learning tasks: supervised learning or unsupervised learning [4].The supervised learning task is to predict some additional aspect of an input object. Examples of such a task are the simple problem of trying to predict a person's weight given their height and the more complex task of trying to predict the topic of an image given the raw pixel values. One core area of supervised learning is the classification task.

Classification is a supervised learning task where the additional aspect of an object that wishes to predict takes discrete values. The additional aspect is the label. The goal in classification is to then create a mapping from input objects to labels. A typical example of a classification task is document categorization, in which it is automatically label a new text document with one of several predetermined topics e.g., sports, politics, business. The machine learning approach to tackling this task is to gather a training set by manually labeling some number of documents. Next it can use a learner together with the labeled training set to generate a mapping from documents to topics. It can be called as a classifier. It can then use the classifier to label new unseen documents.

The other major area of machine learning is the unsupervised learning task. The distinction between supervised and unsupervised learning is not entirely sharp, however the essence of unsupervised learning is that it can not given any concrete information This is in contrast to classification are given manually labeled training data. Unsupervised learning encompasses clustering where groups of data instances that are similar to each other and model building where a model of our domain from our data. One major area of model building in machine learning, and one which is central to statistics is parameter estimation. Here, a statistical model of a domain which contains a number of parameters that need estimating. By collecting a number of data instances we can use a learner to estimate these parameters. Yet another, more recent, area of model building is the discovery of correlation sand causal structure within a domain. The task of causal structure discovery from empirical data is a fundamental problem, central to scientific endeavors in many areas. Gathering experimental data is crucial for accomplishing this task.

For all of these supervised and unsupervised learning tasks, usually we first gather a significant quantity of data that is randomly sampled from the underlying population distribution and then induce a classifier or model. This methodology is called passive Learning. A passive learner is shown in figure 1 receives a random data set from the world and then outputs a classifier or model.

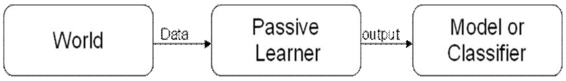


Figure 1: General schema for a Passive Learner

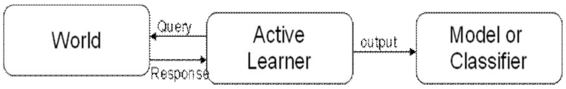


Figure 2: General schema for an Active Learner

1. Active Learners

An active learner is in figure 2 gathers information about the world by asking queries and receiving responses. It then outputs a classifier or model depending upon the task that it is being used for. An active learner differs from a passive learner which simply receives a random data set from the world and then outputs a classifier or model. One analogy is that a standard passive learner is a student that gathers information by sitting and listening to a teacher while an active learner is a student that asks the teacher questions, listens to the answers and asks further questions based upon the teacher's response. It is plausible that this extra ability to adaptively query the world based upon past responses would allow an active learner to perform better than a passive learner, and indeed we shall later demonstrate that, in many situations, this is indeed the case.

2. Pool Based Active Learning

In many supervised learning tasks, labeling instances to create a training set is time consuming and costly; thus, finding ways to minimize the number of labeled instances is beneficial. Usually, the training set is chosen to be a random sampling of instances. However, in many cases active learning can be employed. Here, the learner can actively choose the training data. It is hoped that allowing the learner this extra flexibility will reduce the learner's need for large quantities of labeled data.

Pool-based active learning for classification was introduced by Lewis and Gale (1994). The learner has access to a pool of unlabeled data and can request the true class label for a certain number of instances in the pool. In many domains this is a reasonable approach since a large quantity of unlabeled data is readily available. The main issue with active learning is finding a way to choose good requests or queries from the pool. Examples of situations in which pool-based active learning can be employed are:

• *Web search:* A Web-based company wishes to search the web for particular types of pages e.g., pages containing lists of journal publications. It employs a number of people to hand-label some web pages so as to create a training set for an automatic classifier that will eventually be used to classify the rest of the web. Since human expertise is a limited resource, the company wishes to reduce the number of pages the employees have to label. Rather than labeling pages randomly drawn from the web, the computer requests targeted pages that it believes will be most informative to label.

• *Email filtering:* The user wishes to create a personalized automatic junk email filter. In the learning phase the automatic learner has access to the users past email files. It interactively brings up past email and asks the user whether the displayed email is junk mail or not. Based on the user's answer it brings up another email and queries the user. The process is repeated some number of times and the result is an email filter tailored to that specific person.

• *Relevance feedback*: The user wishes to sort through a database or website for items such as images, articles, etc. that are of personal interest—an "I'll know it when I see it" type of search. The computer displays an item and the user tells the learner whether the item is interesting or not. Based on the user's answer, the learner brings up another item from the database. After some number of queries the learner then returns a number of items in the database that it believes will be of interest to the user.

II. RELATED WORK

Involving domain experts into learning is a common and often necessary step in domain-driven data mining [3]. It has been shown that, in many real-world applications, domain experts can play an important role in the entire knowledge discovery and data mining process [5][6]. Specifically, the learning algorithms guided by domain experts can achieve significantly better performance than the automated data only learning. Thus, active learning has been intensively studied, due to its natural capability of integrating domain experts into the learning process.

Most previous works of active learning can be divided into two paradigms: the pool-based active learning and the membership query. In the pool-based active learning, a pool of unlabeled examples is given, and the learner can only choose examples to label from the pool [7]. Briefly speaking, the pool-based active leaner first evaluates each example in the pool, to decide which one can at maximum improve the performance of the current model. Then, the learner acquires its label from oracle to update the labeled training set and the learning model, and the process repeats.

On the other hand, active learning with membership queries (or direct query construction) can construct examples (without the need of the pool) and request labels. Both of these active learning methods reduce the number of labeled examples needed, compared with labeling examples randomly. The essence of active learning lies in the "goodness" measurement of the unlabeled examples with respect to the current model. Many criteria have been proposed in the literatures. Uncertainty sampling considers the most uncertain example as the most valuable one, and has been thoroughly studied and widely used in many previous researches. Query-by-committee [8] is a more theory-based approach, and considers the example minimizing the version space as optimal.

Besides, other criteria, such as variance reduction, Fisher information ratio, and estimated error reduction, are also elaborately designed and well accepted in active learning research area. In this paper, the proposed AGQ algorithm can be integrated with any of the above criteria, and the most widely used uncertain sampling is chosen for illustration.

All previous works of active learning assume that the oracle could only answer specific queries, with all attribute values provided. To the best of our knowledge, our AGQ algorithm in this project is the first work of active learning with generalized queries. Again, the main advantage of AGQ is that one generalized query is usually equivalent to many specific ones. Thus, the answer from the oracle is also for all of the specific queries. Even though one generalized query is equivalent to multiple specific queries, our AGQ method is still quite different from batch-mode active learning [9]. In batch-mode active learning, the learning model requests labels for a batch of examples (i.e., multiple specific queries) in each iteration; thus, the oracle is required to provide multiple answers for all these queries (i.e., with multiple costs). On the other hand, in AGQ, the oracle answers only one generalized query in each iteration. Thus, AGQ costs much less than the batch-mode active learning, for answering queries in the learning process. Druck et al [10] proposed active learning with feature labeling, which queries the label for one specific feature and is mainly used in natural language processing.

Although feature labeling is considered similar to the generalized query, AGQ algorithm is significantly different in the following three aspects: First, instead of querying label for one specific feature, our AGQ could query the labels for multi feature. Thus, feature labeling is essentially a special case of our AGQ. In other words, our generalized query is a generic paradigm for both instance-based queries and feature-based queries. Second, AGQ always finds the most uncertain example (when integrated with uncertain sampling) and generalizes it to a query. Labeling such uncertain examples has been proved to be very effective in improving predictive accuracy. On the other hand, feature labeling generally finds the most predictive or most frequent)feature for querying; thus, the answer from the oracle may not provide much new information to improve the model. Third, and most importantly, as feature labeling always queries label for only one feature, the answer from the oracle could be very uncertain.

III. MOTIVATION

The main aim of the paper is to have a generalized query in different manner using AGQ algorithm. The answer of the queries is equivalent to many specific once that means it gives multiple answers. The present defines that the information’s are extracted from the domain experts. In that existing system it can get the information only for the specific queries. But in this, it can get the information about the generalized queries with multiple answers.

In the existing system gives the information only for specific queries. It does not give multiple options for the queries. Learner wants to label the database or pool and Indexing method is not introduced. There are many issues in the proposed system. It gives the information for specific queries as well as generalized queries , gives the multiple answers for one generalized queries that means one generalized query is usually equivalent to many specific ones and Labeling and indexing were introduced.

IV. REQUIREMENT ANALYSIS AND SPECIFICATION

The purpose of the Software Requirement Specification is to produce the specification of the analysis task and also to establish complete information about the requirement, behavior and other constraints such as functional performance and so on. The goal of Software Requirement Specification is to completely specify the technical requirements for the software product in a concise and unambiguous manner. it has been implemented in JSP and MySQl. MYSQL is a Structure Query Language used to create and manage computer-based databases on desktop computers and/or on connected computers.

V. AGQ ALGORITHM

Domain-driven data mining actively involves human experts in the learning process. Active learning naturally puts human experts in the process. In this paper, a new active learning paradigm has been designed in which the learner can ask generalized queries, and assume that the domain expert or oracle can answer such generalized queries.

In this paper, AGQ can generalize attributes (nominal or numeric) with specific values. An initial learner L is built using the current labeled training data set R. Then, L is used to predict each example in the pool U. The most uncertain example from the Pool is chosen.

AGQ, then, finds irrelevant attributes in the most uncertain example. AGQ submits this generalized query to the oracle, which will return a label. AGQ will utilize the label and then update the training data, and iterate to Step 1 ,that is to continue learning actively.

VI. MODULES

In the asking generalize queries has four modules such as User Query Process, Extracting the keywords using the AGQ+, Active Learner, Labeling and Indexing the database.

A. User Query Process

User must need an authorization. Authorization is for only prescribed users entering the querying scheme rather than unauthorized access. Users and Server have an authorization entry. Sometimes by mistake, the user gives wrong user name and password, the server generates the warning to every mismatch inputs. If it is matched, then the user get the connection otherwise the server quits the unauthorized person. After giving the user name and password matched user can generate the generalized query.

B. Extracting The Keywords Using The AGQ

In the algorithm from the keyword it will map the irrelevant and the relevant attributes and from the relevant attributes we are going to extract the two concepts named as a nominal attributes and the numerical attributes. Irrelevant attributes are marked as a "*". For example, to predict "osteoarthritis," "knee pain" could be a relevant nominal attribute with values "none," moderate," and "severe," and "age" could be another relevant attribute with numeric values. Then, in addition generalizing the irrelevant attributes as "*," it may also generalize the relevant attributes to several nominal values such as, "knee pain" being "moderate" or "severe") or numeric interval (such as, "age" being 50, 65). It can, then, construct generalized queries, such as "are people aged between 50 and 65, with moderate or severe knee pain, likely to have osteoarthritis?".

C. Active Learner

Domain-driven data mining actively involves human experts in the learning process. Active learning naturally puts human experts in the process. In this project, a new active learning paradigm is used in which the learner can ask generalized queries, and it is assumed that the domain expert or oracle can answer such generalized queries. In this AGQ to consider the database as the pool and clustering the pool files presented in the database and giving the label for the each cluster this labeling mechanism is done by the active learner. In the given label, the database can be indexed for the fast retrieval of the data from the database. This makes the list that is presented in the database.

D. Indexing The Database

In the given label, index the detail for the fast retrieval of the data from the database. This makes the list that is presented in the database.

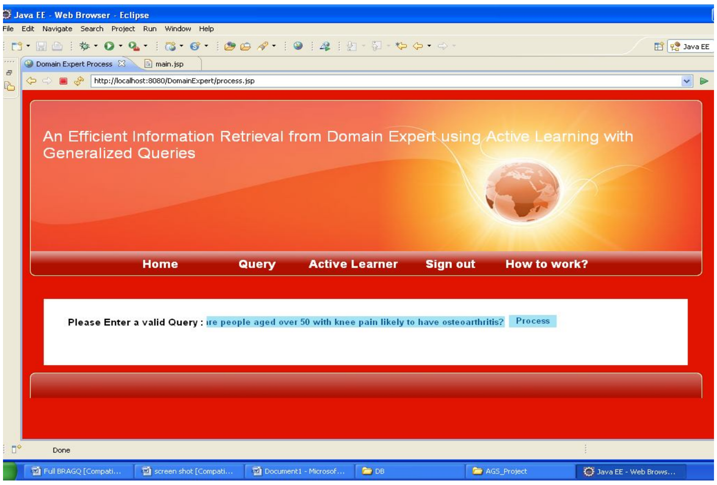
VII.PERFORMANCE ANALYSIS

In recent years, domain-driven data mining (D3M) has received extensive attention in data mining. Unlike the traditional data-driven data mining, D3M tends to discover actionable knowledge by tightly integrating the data mining methods with the domain-specific business processes. However, in most cases, the domain specific actionable knowledge cannot be discovered without the support of domain knowledge, mainly provided by human experts. Thus, the human-machine-cooperated interactive knowledge discovery process is widely applied in real-world applications. Motivated by domain-driven data mining, in this paper, we attempt to maximize the utility of domain experts or oracles in active learning process.

Specifically, traditional active learning algorithms only assume that the domain expert or oracle is capable of answering specific queries is in figure 3, with all attribute value provided. For example, if the task is to predict osteoarthritis based on a patient data set with 30 attributes, the previous active learners could only ask the specific queries as: does this patient have osteoarthritis, if ID is 32765, name is Jane, age is 35, gender is female, weight is 85 kg, and blood pressure is 160/90, temperature is 98F, no pain in the knees, no history of diabetes, and so on (for all 30 attributes). Many of these 30 attributes may not be relevant to osteoarthritis in this case. Not only could specific queries like this confuse the domain experts (oracles), but the answers returned are also specific: each label given is only for one specific query.

In this paper, it is assumed that the oracle is capable of answering generalized queries, and then a novel active learning paradigm in which such generalized queries can be asked and answered. An algorithm called AGQ, for Active learner with Generalized Queries, can construct generalized queries with don't-care attributes, for either the pool-based or the membership query active learner. However, AGQ can only generalize specific attribute values to don't care. Then, AGQ can generalize specific attribute values to meaningful new features for both nominal and numeric attributes.

For example, AGQ can ask such queries as "are people aged between 50 and 65, with moderate or severe knee pain, likely to have osteoarthritis?" Here, age (a numeric attribute) is generalized to a range, and knee pain (a nominal attribute) is generalized to a subset of values. These newly constructed features can form hierarchical structures, and are often meaningful in real-world applications.



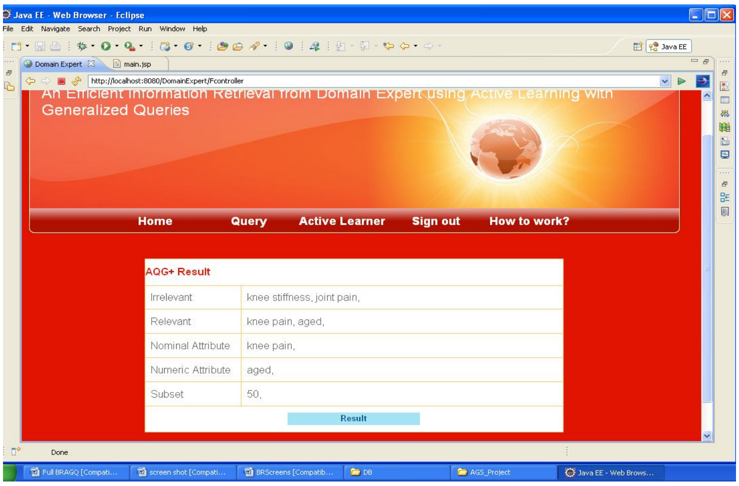
http://localhost:8080/DomainExpert/process.jsp

An Efficient Information Retrieval from Domain Expert using Active Learning with Generalized Queries

Please Enter a valid Query : ıre people aged over 50 with knee pain likely to have osteoarthritis?

Figure 3. User Query Form

User must need an authorization. Authorization is for only prescribed users entering the querying scheme rather than unauthorized access. Users and Server have an authorization entry. Sometimes by mistake, the user gives wrong user name and password, the server generates the warning to every mismatch inputs. If it is matched, then the user get the connection otherwise the server quits the unauthorized person. After giving the user name and password, control goes to the Active Learner screen. To Train the system give the detail to the database. After giving the user name and password matched user can generate the generalized query. After giving the user name and password matched user can generate the generalized query. AGQ can also automatically produce subsets for nominal attributes and ranges for numeric attributes. The Result figure 4 gives the information available in the database.



http://localhost:8080/DomainExpert/Fcontroller

An Efficient Information Retrieval from Domain Expert using Active Learning with Generalized Queries

AQG+ Result

Irrelevant	knee stiffness, joint pain,
Relevant	knee pain, aged,
Nominal Attribute	knee pain,
Numeric Attribute	aged,
Subset	50,

Figure 4. Result Form

VIII. CONCLUSION

Domain-driven data mining calls for domain experts Involvement in the data mining process. Active learning involves domain experts in its need to obtain label information for the queries. However, previous active learning algorithms assume that the oracle can only answer specific queries that represent single examples. However, in real-world applications, the domain experts are often more readily to answer “generalized queries” with don’t-care attributes and generalized. Answers to such generalized queries can provide more information to improve learning.

In this paper, the active learning paradigm in which the learner can ask generalized queries, and the domain expert or oracle can answer such generalized queries. AGQ can construct generalized queries with don’t-care attributes, for either the pool-based or the membership query active learner. Then, AGQ can generalize specific attribute values to meaningful new features for both nominal and numeric attributes. These newly constructed features can form hierarchical structures, and are often meaningful in real-world applications.

Experiments on sample data sets show that AGQ ask significantly fewer queries compared with the traditional active leaner. In addition, AGQ can also automatically produce subsets for nominal attributes and ranges for numeric attributes, which can be used in further learning. To the best of my knowledge, this active learning with generalized queries (AGQ) gives effective result.

IX. FUTURE ENHANCEMENT

The difficulty of generalized queries is that the answers from the oracle can be uncertain, thus noisy labels might be introduced and performance might be degraded. This easily happens especially when the initial labeled training set is small. Strategies for dealing with highly uncertain answers from the oracle, and for preventing dramatic changes of data distribution when new examples are included in the training set are also interesting issues to further improve the performance of AGQ.

REFERENCES

[1] Jun Du, Charles X. Ling, Asking Generalized Queries to Domain Experts to Improve Learning, IEEE Transaction on Knowledge and Data Engineering, Vol.22, No,16, June 2010.
[2] J. Du and C. X. Ling. Active learning with generalized queries. In Proceedings of the 2009 IEEE International Conference on Data Mining, 2009.
[3] L. Cao and C. Zhang, Domain-driven data mining: A practical methodology. International Journal of Data Warehousing and Mining, 2(4):49–65, 2006.
[4] Simon Tong, Active Learning: Theory and Applications, Pg 2-5.
[5] C. C. Aggarwal. Towards effective and interpretable data mining by visual interaction. In SIGKDD Explorations, volume 3, pages 11–22, 2002.
[6] M. Ankerst. Report on the sigkdd-2002 panel the perfect data mining tool: interactive or automated? SIGKDD Explorer. Newsl. 4(2):110–111, 2002.
[7] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In W. W. Cohen and H. Hirsh, editors, Proceedings of ICML-94, 11th International Conference on Machine Learning, pages 148–156, New Brunswick, US, 1994. Morgan Kaufmann Publishers, San Francisco, US.
[8] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee.In COLT ’92: Proceedings of the fifth annual workshop on Computational learning theory, pages 287–294, New York, NY, USA, 1992. ACM Press.
[9] S. C. H. Hoi, R. Jin, J. Zhu, and M. R. Lyu. Batch mode active learning and its application to medical image classification. In ICML ’06: Proceedings of the 23rd international conference on Machine learning, pages 417–424, New York, NY, USA, 2006. ACM.
[10]G. Druck, G. S. Mann, and A. McCallum. Learning from labeled features using generalized expectation criteria. In S. H. Myaeng, D. W. Oard, F. Sebastiani, T. S. Chua, M. K. Leong, S. H. Myaeng,D. W. Oard, F. Sebastiani, T. S. Chua, and M. K. Leong, editors,SIGIR, pages 595–602. ACM, 2008.
[11].B. Liu, W. Hsu, and Y. Ma. Integrating classification and association rule mining. In Knowledge Discovery and Data Mining, pages 80–86, 1998.
[12]O. Longbing Ca and G. Chengqi Zhan. The evolution of kdd: Towards domain-driven data mining. International Journal of Pattern Recognition and Articial Intelligence (IJPRAI), 21(4):677– 692, 2007.
[13]C. X. Ling and J. Du. Active learning with direct query construction. In KDD ’08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 480–487, New York, NY, USA, 2008. ACM.