INTERNATIONAL JOURNAL OF INNOVATIVE TECHNOLOGY AND CREATIVE ENGINEERING (ISSN:2045-8711)
VOL.5 NO.3 MARCH 2015

A Domain Ontology Method for Semantic Conceptual Distance based on Rule Ranking Algorithm

S.Antoinette Aroul Jeyanthi
Ph.D (Research Scholar), Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India.
Email: jayanthijames@yahoo.com
Dr.S.Pannirselvam
Research Supervisor & Head
Department of Computer Science, Erode Arts & Science College (Autonomous), Erode, Tamil Nadu, India.
Email: pannirselvam08@gmail.com

Abstract— The problem of finding interesting and actionable patterns is a major challenge in data mining. It has been studied by many data mining researchers. The issue is that data mining algorithms often generate too many patterns, which make it very hard for the user to find those truly useful ones. Evaluating and ranking the interestingness or usefulness of association rules is important in data mining. In this paper, proposed and implemented an approach for ranking the rules using the semantic conceptual distance based on domain ontology which is represented as DAG.

Keywords— Domain ontology, Unexpectedness, Association rule, Interestingness, Conceptual distance.

1. INTRODUCTION

Knowledge discovery in data mining has been defined in Fayyad et al., [1] as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns from data. Association rule algorithms Agrawal et al., [2] are rule-discovery methods that discover patterns in the form of IF-THEN rules. It has been noticed that most of the algorithms that perform data mining generate a large number of rules that are valid but obvious or not interesting to the user. To address this issue, most of the approaches to knowledge discovery use objective measures of interestingness for the evaluation of the discovered rules, such as confidence and support measures. These approaches capture the statistical strength of a pattern. The interestingness of a rule is essentially subjective. Subjective measures of interestingness, such as unexpectedness. Assume that the interestingness of a pattern depends on the decision-maker and does not solely depend on the statistical strength of the pattern. Although objective measures are useful, they are insufficient in the determination of the interestingness of the rules. One way to address this problem is by focusing on discovering unexpected patterns, where the unexpectedness of the discovered patterns is usually defined relative to a system of prior expectations.

Moreover, ontology represents knowledge. Ontology is organized as a DAG (Directed Acyclic Graph) hierarchy. Ontologies allow domain knowledge to be represented explicitly and formally in such a way that it can be shared among human and computer systems. In this paper, we propose a new approach that adds intelligence and autonomy for ranking rules according to their conceptual distance (the distance between the antecedent and the consequent of the rule) relative to the hierarchy. In other words, highly related concepts are grouped together in the hierarchy. The more distant the concepts are, the less they are related to each other. For concepts that are part of the definition of a rule, the less the concepts are related to each other, the more the rule is surprising and therefore interesting. With such a ranking method, a user can check fewer rules on the top of the list to extract the most pertinent ones.

2. LITERATURE SURVEY

The unexpectedness of patterns has been studied in defined in comparison with user beliefs. A rule is considered to be interesting if it affects the levels of conviction of the user. The unexpectedness is defined as a distance, and it is based on a syntactic comparison between a rule and a conviction.

Padmanabhan [3] et al., focused is on discovering minimal unexpected patterns rather than using any of the post-processing approaches, such as filtering, to determine the minimal unexpected patterns from the set of all of the discovered patterns.

Liu [4] et al., developed a subjective interestingness (unexpectedness) of a discovered pattern is characterized by asking the user to specify a set of patterns according to his/her previous knowledge or intuitive feelings. This specified set of patterns is then used by a fuzzy matching algorithm to match and rank the discovered patterns.

Sahar [5] studied a genetic algorithm to dynamically maintain and search populations of rule sets for the most interesting rules rather than act as a post-processor.

McGarry [6] identified by the genetic algorithm compared favorably with the rules selected by the domain expert. To find subjectively interesting rules, most existing approaches ask the user to explicitly specify what types of rules are interesting and uninteresting, then generate or retrieve those matching rules. This research on the unexpectedness makes a syntactic or semantic comparison between a rule and a belief.

3. METHODOLOGY

3.1 RULE INTERESTINGNESS MEASURES

In this research work in data mining has shown that the interestingness of a rule can be measured using objective measures and subjective measures. Objective measures

involve analyzing the rule's structure, predictive performance, and statistical significance. In association to rule mining, such measures include support and confidence. However, it is noted in that such objective measures are insufficient for determining the interestingness of a discovered rule. Indeed, subjective measures are needed. There are two main subjective interestingness measures, namely unexpectedness and action ability.

Unexpectedness: Rules are interesting if they are unknown to the user or contradict the user's existing knowledge.

Actionability: Rules are interesting if the user can do something with them to his/her advantage.

3.2 CONCEPTUAL DISTANCE

Two main categories of algorithms for computing the semantic distance between terms organized in a hierarchical structure have been proposed in the literature: distance-based approaches and information content-based approaches. The general idea behind the distance-based algorithms is to find the shortest path between two concepts in terms of the number of edges. The shorter the path from one node to the other, the more similar they are. The problem with this approach is that it relies on the notion that edges in taxonomy represent uniform distances. Information content-based approaches are inspired by the perception that pairs of concepts that share many common contexts are semantically related. The more information that two concepts share in common, the more similar.

The problem of the ontology distance is that it is highly dependent on the construction of the ontology. The measure is, highly dependent on oftentimes subjective ontology engineering decisions. To address this problem, we are associating a weight to any concept in the ontology that represents the degree of importance of this concept in the ontology along with the strength of any relation between the concepts. In an IS-A semantic network, the simplest form of determining the distance between two concept nodes, A and B, is the shortest path that links A and B. The minimum number of edges that separate A and B or the sum of the weights of the arcs along the shortest path between A and B.

3.3 PROPOSED METHODOLOGY

The feature extraction is an important process to make efficient rule ranking algorithm. Hence an enhanced technique is used to extract the feature.

3.3.1 ONTOLOGY BASED RULE RANKING ALGORITHM

The technique analyzes rules and detects the interrelation between various diseases and symptoms which are not directly associated in the dataset. The weight was set to 1 for all the symptoms that are directly associated to diseases. For the indirect symptoms the weight was increased by 1 at each level. Then the algorithm computes their semantic conceptual distance. The larger the distance is, the more the rule is interesting.

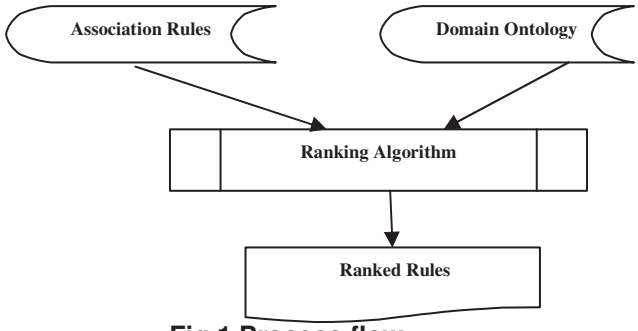

Fig.1 Process flow

A. Concept Semantic Distance

The semantic distance between the two concepts A and B is the sum of the weights of the arcs along the shortest path between A and B. To compute the shortest path between two nodes using Dijkstra's algorithm. To compute the distance between groups of concepts, for a given rule R: $X \rightarrow Y$, where $X = X1 \wedge \ldots \wedge Xk$, $Y = Y1 \wedge \ldots \wedge Ym$, using the Hausdorff distance. The function h(X,Y) is called the directed Hausdorff distance from X to Y. This expression measures the conceptual distance between groups $X = X1 \wedge \ldots \wedge Xk$ and $Y = Y1 \wedge \ldots \wedge Ym$ of the concepts that contain the k Xi and m atomic Yj concepts, respectively.

B. Rule Ranking Algorithm

In this section, introduce an algorithm to rank the rules according to their conceptual distance based on a domain ontology that represents the background knowledge. The rules that we consider are in the form of "body → head", where "body" and "head" are conjunctions of concepts in the vocabulary of the ontology. Here assume that other techniques carry out the task of pattern discovery and eliminate the patterns that do not satisfy the objective criteria. With such a ranking, a user can check only confirm the rules that are the most pertinent.

3.4 PROPOSED ALGORITHM

The entire procedure is presented as simple algorithms.

3.4.1 Algorithm – I

ND: Number of nodes
R: Set of rules
R = {Ri/Ri = body → head} where i ∈ [1,N]
N: number of rules
D: Maximum depth of the hierarchy
Xi, Yj: Atomic Concepts; i ∈ [1,k]; j ∈ [1,m]
Body = X1 ∧.. ∧Xk
Head = Y1 ∧.. ∧Ym
for i = 1 to ND
for j = 1 to ND
Begin
// ShortestPath(Xi,Xj) shortest path between Xi and Xj//
//Make a call to the weight(ShortestPath(Xi,Xj) above//
Distance(Xi,Xj) = weight(ShortestPath(Xi,Xj));
End
for i = 1 to N
Distance(Ri) = (Distance(X1 ∧.. ∧Xk; Y1 ∧.. ∧Ym));
Sort Distance(Ri) descending;

4. EXPERIMENTATION & RESULTS

In this experiment, the specific set of discovered rules is considered as input for the algorithm. The dynamic formation of the associations and interconnection between diseases and respective symptoms are implemented. Using dynamic formation of the graph data structure, the distance based on the depth based search is possible. At the initial level, the distance from the graph data structure of association and dependent rules is evaluated for all the instances. The following set of rules are considered as input for our experiment

Rule1: *headache, cold, cough -> fever*
Rule2: *hairloss, weightloss, swallow-> cancer*
Rule3: *sneeze, cough, cold-> viral*
Rule4: *cough, fever, fatique->flu*
Rule5: *fever, hairloss, fatique-> cancer*

The domain ontology which is represented as DAG of the given set of rules is as follows

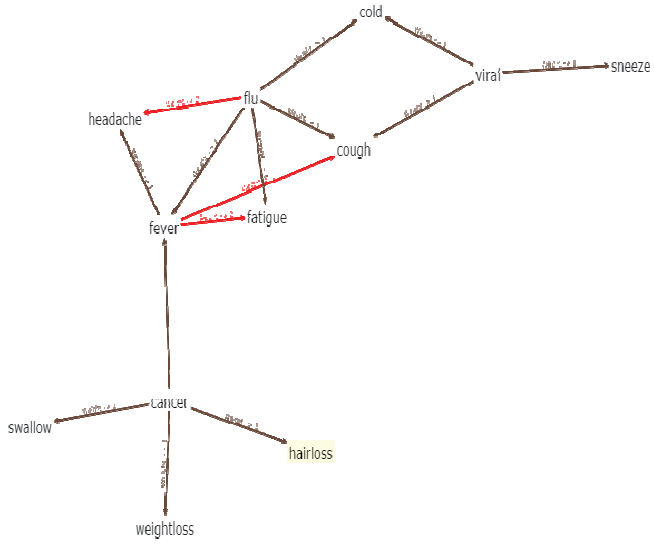

Fig.2 Set of Rules

Using the proposed implementation, the hidden association or interrelation between assorted aspects are fetched and evaluated. The distance between the various symptoms is calculated.

dist(hairloss,sneeze) = 5

dist(weightloss,sneeze) = 5

dist(swallow,sneeze) = 5

dist(fatique,sneeze) = 4

dist(hairloss,headache) = 3

dist(weightloss,headache) = 3

dist(swallow,headache) = 3

dist (fatique,headache)= 3

dist (sneeze, fever) -> 3

dist (headache, viral) -> 3

From the results, the rules with the highest distance are considered as unexpected rules and are given the highest rank. The symptoms associated with that rule do not have direct association. They are indirectly associated with each other. The decision maker has to give more attention to the rare combination of symptoms than the common, directly associated symptoms. In the other ranking algorithms, the directly associated symptoms have the highest rank.

7. CONCLUSION

In this paper, proposed a new approach for ranking association rules according to their conceptual distance, which was defined on the basis of the ontological distance. The proposed ranking algorithm helps the user to identify interesting association rules, particularly indirectly associated and unexpected rules. This algorithm uses domain ontology to calculate the distance between the antecedent and the consequent of the rules on which the ranking is based. The larger the conceptual distance is, the more the rule represents a high degree of interest.

8. REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth,From data mining to knowledge discovery in databases,AI Mag., 17 (3) (1996), pp. 37–54.

[2] Agrawal, R., Imieliński, T., Swami, A., 1993. Mining association rules between sets of items in large databases", ACM SIGMOD Record, vol. 22(2), pp. 207–216.

[3] B.Padmanabhan, A. Tuzhilin,On characterization and discovery of minimal unexpected patterns in rule discovery IEEE Trans. Knowl. Data Eng., 18 (2) (2006), pp. 202–216.

[4] B.Liu, W.Hsu, S.Chen, Y.Ma,Analyzing the subjective interestingness of association rules, Intelligent Sys. Appl. IEEE,15 (5) (2000), pp.47–55.

[5] Sahar,S., 2002. On incorporating subjective interestingness into the mining process, In Data Mining, 2002. ICDM 2003. Proceedings. Of the 2002 IEEE International Conference on Data Mining, pp. 681–684.

[6] K. McGarry,A survey of interestingness measures for knowledge discovery Knowl. Eng. Rev., 20 (01) (2005), pp. 39–61.

[7] R.Agrawal, T.Imielinski, A.Swami,Database mining: A performance perspective Knowl. Data Eng. IEEE Trans., 5 (6) (1993), pp. 914–925.

[8] R.Agrawal, R.Srikant, Fast algorithms for mining association rules, Proceedings of the 20th International Conference Very Large Data Bases, 1215 (1994), pp. 487–499.

[9] Farzanyar, Z., kangavari, M., Hashemi, S., 2006. A new algorithm for mining fuzzy association rules in the large databases based on ontology, Proceedings of the Sixth IEEE International Conference on Data Mining, pp. 65–69.

[10] B. Padmanabhan, A. Tuzhilin,Unexpectedness as a measure of interestingness in knowledge discovery Decision Support Sys., 27 (3) (1999), pp. 303–318.