جـامـعـة الـبـعث
كلّيـة العلـوم
قسـم الرّيـاضيات – معلوماتية

استخدام نظرية المجموعات التقريبية لاكتشاف المعرفة من قواعد المعطيات

إعداد الطالب
أنـس المصطفى

٢٠١١ م – ١٤٣٢ هـ

بإشراف
د. ناصر أبو صالح د. عبد الله العمر

ملخص

يعتبر اكتشاف المعرفة من قواعد المعطيات منطقة بحثية نشطة تبشر بربح مرتفع في العديد من تطبيقات الأعمال والتطبيقات العلمية. وإن التحدي الرئيسي لاكتشاف المعرفة من قواعد المعطيات هو المعالجة الآلية لكميات المعطيات الخام والكبيرة، وتحديد الأنماط الأكثر أهميةً والتي لها معنى، وتقديم هذه المعرفة في شكل مناسب لتحقيق هدف المستخدم. وتواجه نظم اكتشاف المعرفة مشاكل تحدٍ من قواعد معطيات العالم الحقيقي التي تتجه لأن تكون كبيرة جداً، ومكررة، ومفعمة بالضجيج، وديناميكية. ولقد تم تناول كل مشكلة من هذه المشاكل إلى مدى معين ضمن تعلم الآلة. معالجة كل هذه المشاكل أثناء إنتاج معرفة مفيدة بشكل فعال ومؤثر هي المركز الرئيسي لهذه الأطروحة. ومن أجل معالجة كل هذه المشاكل كان هدفنا الرئيسي في هذه الأطروحة هو إبراز العلاقة والارتباط بين نظرية المجموعات التقريبية واكتشاف المعرفة من قواعد المعطيات، والتحقق من كيفية استغلال هذه النظرية في اكتشاف المعرفة من قواعد المعطيات، وكذلك تقييم الفائدة المحتملة من القواعد المكتشفة. وأهمية البحث في نظرية المجموعات التقريبية تعود إلى حقيقة أنها تستخدم فقط المعلومات الداخلية لا تعتمد على افتراض نموذج إضافي كما تفعل الطرق الأخرى. وهذا يعني أنه بدلاً من استخدام أعداد أو وسطاء إضافية أخرى، يستفيد تحليل المجموعة التقريبية فقط من بنية المعطيات المعطاة. وسمة هامة أخرى لهذه النظرية هي قدرتها على التعامل مع الضجيج والغموض وعدم التأكد والمعطيات غير المتسقة حيث أنها تعالج بنفس طريقة معالجة المعطيات المتسقة ولا تتطلب أي طرق إضافية.

في إنجاز هدفنا، تتألف منهجية البحث المتبعة في هذه الأطروحة من عدة خطوات. في الخطوة الأولى، عرضنا مراجعة لعملية اكتشاف المعرفة من قواعد المعطيات، حيث في هذه المرحلة راجعنا مفهوم التعلم من الأمثلة ومكوناته الأساسية، وقدمنا شرح مبسط عن بعض خوارزميات التعلم من الأمثلة المطورة، وأخيراً قدمنا تلخيص عن بعض نماذج التعلم من قواعد المعطيات. وفي الخطوة الثانية، عرضنا مقدمة عن نظرية المجموعات التقريبية. حيث في هذه المرحلة، تمت مناقشة الفلسفة الأساسية لهذه النظرية، وإظهار الفروقات الرئيسية بين نظرية المجموعات التقليدية ونظرية المجموعات التقريبية، ومناقشة علاقة نظرية المجموعات التقريبية

بالنظريات الأخرى التي تتعامل مع الغموض وعدم التأكد، وأخيراً مناقشة المفاهيم الرئيسية للنظرية وأهميتها في التعامل مع الغموض وعدم التأكد. وفي الخطوة الثالثة، صممنا إطار عملنا المقترح لاكتشاف المعرفة من قواعد المعطيات اعتماداً على نظرية المجموعات التقريبية. حيث في هذه المرحلة، تم اقتراح وتطوير عدة خوارزميات لتوليد القواعد المحددة وغير المحددة والقواعد المؤكدة، والقواعد الممكنة، وخوارزمية لتوليد القاعدة الأفضل وخوارزمية للتصنيف الأفضل. وفي الخطوة الرابعة، تم بناء نظام تعلم قاعدة معطيات تجريبي لاختبار فعالية الخوارزميات المطورة مقابل معطيات العالم الحقيقي.

Abstract

The knowledge Discovery in Databases (KDD) is an active research area with the promise for a high profit in many business and scientific applications. The grand challenge of knowledge discovery in databases is to automatically process large quantities of raw data, identify the most significant and meaningful patterns, and present this knowledge in an appropriate form for achieving the user's goal. The knowledge discovery systems face challenging problems from the real-world databases which tend to be very large redundant, noisy, and dynamic. each of these problems has been addressed to some extent within the machine learning, but few systems address them all, Collectively handling these problems while producing useful knowledge efficiently and effectively is the main focus of this thesis. In this thesis, for handling these problems, the main objective in our study is to highlight the relationship and the connection between the rough sets theory and field of knowledge discovery in databases, and check of how to exploit this theory in knowledge discovery from databases, as well as evaluating the potential usefulness of the discovered rules. The research importance of in rough set theory resides in the fact that it uses only internal information and does not rely on additional model assumption as other methods do that is, instead of using numbers or other additional parameters, rough set analysis utilized solely the structure of the given data. Another important feature of this theory is its ability on handling with noisy and inconsistent, ambiguity, uncertainty and conflicting data where they are handled with the same way as handling consistent data and they does not require any additional methods. In achieving our main objective, the research methodology followed in this thesis consists of several steps. In the first step, we presented an overview of the knowledge discovery in databases, where, in this step, we presented some theoretical issues related to learning from examples, and recent progress in knowledge discovery in database systems and knowledge base systems which adopt the learning from examples paradigm. In the second step, we presented an introduction to the rough set theory. Where at this stage, was discussed the basic philosophy of this theory, and show major differences between the traditional set theory and the rough set theory, and discuss the

relationship rough set theory with other theories dealing with ambiguity and uncertainty, and finally discuss the mainly concepts of the theory and its importance in dealing with ambiguity and uncertainty. In the third step, we presented our proposed framework for rough set based knowledge discovery. In the fourth step, we presented implementation and experiment for testing the effectiveness of some learning algorithms against real world data sets.