# BAB I

# PENDAHULUAN

## 1.1 Latar Belakang Masalah

*Data mining* merupakan sebuah proses untuk menggali informasi yang tersembunyi atau menemukan pola yang bermanfaat dalam sekumpulan data yang besar. Untuk mendukung proses penggalian informasi ini terdapat metode-metode dalam *data mining* yaitu asosiasi, *clustering* dan klasifikasi. *Clustering* merupakan salah satu teknik yang digunakan dalam *data mining* yang bertujuan untuk mengelompokkan suatu objek yang memiliki *similarity* tinggi dalam satu *cluster* dan yang memiliki *dissimilarity* tinggi di *cluster* yang berbeda (Anil dan Richard, 1998). Data yang digunakan dalam *clustering* dapat berupa data numerik dan kategori. Saat ini pada kategori data merupakan topik yang sedang banyak dilakukan dalam penelitian (Jain, A.K, 1998).

Data kategori adalah data yang diambil dari suatu himpunan nilai tertentu yang tidak harus berupa angka atau pecahan, misalnya [*male,female*], [*low,medium,high*]. Data kategori tidak dapat diukur dan diurutkan karena tidak dapat dibandingkan mana yang memiliki nilai yang lebih besar atau yang lebih kecil. Pada *clustering*, data kategori tidak memungkinkan menggunakan konsep geometri (Z Huang, 1998). Huang mengusulkan algoritma K-Modes untuk mengatasi masalah pada data bersifat kategori yang tidak dapat dihitung menggunakan konsep geometri.

Proses algoritma K-Modes dimulai dengan membentuk terlebih dahulu partisi *cluster* dan menentukan titik pusat di awal proses kemudian secara iteratif. Partisi *cluster* ini terus diperbaiki hingga tidak ada lagi perubahan yang terjadi pada setiap partisi *cluster*. Dilihat dari prosesnya, algoritma K-Modes tergolong algoritma *data mining* yang cepat dan mudah diimplementasikan, Walaupun demikian terdapat beberapa kekurangan yang dimiliki K-Modes salah satunya adalah penentuan titik pusat awal (*initial centroid*) karena penentuan titik pusat awal yang dilakukan secara acak.

Selain sensitif pada penentuan *centroid* awal, algoritma K-Modes juga sensitif terhadap penentuan pengukuran jarak *cluster*. K-Modes menggunakan ukuran *matching* (cocok) atau *mismatching* (tidak cocok) untuk menemukan kesamaan antar objek. *Hamming Distance* atau *Simple Matching Measure* adalah teknik yang digunakan untuk menemukan *similarity* antara dua objek yang akan ditempatkan pada satu *cluster* pada algoritma K-Modes. Namun pada perhitungan ukuran ini tidak mempertimbangkan hubungan kesamaan yang ada pada nilai atribut sehingga menghasilkan kesamaan *intra cluster* yang lemah.

Untuk menemukan kesamaan antar objek pada tipe data kategori, pada penelitian sebelumnya telah memodifikasi algoritma K-Modes untuk memperbaiki dalam pengukuran jarak *cluster* menggunakan *Weighted Dissimilarity Measure* (WDM) sehingga menghasilkan *cluster* lebih akurat dari K-Modes tradisional (Aranganayagi dan Thangavel, 2009). Penelitian sebelumnya telah mengimplementasikan algoritma pada *dataset* dari berjumlah kecil hingga

besar. *Dataset* merupakan himpunan data yang sudah teruji kebenarannya dan sering digunakan pada beberapa penelitian di wilayah *data mining*.

Akan tetapi algoritma WDM K-Modes masih mengadopsi penentuan *centroid* awal yang sama seperti K-Modes, dimana penentuan *centroid* awalnya ditentukan secara acak. Apabila *centroid* awal yang ditentukan sistem tepat maka hasil *cluster* dari WDM K-Modes dapat optimal, sebaliknya jika penentuan *centroid* awal pada WDM K-Modes tidak tepat maka hasil *cluster*nya pun akan kurang optimal. Sehingga dibutuhkan algoritma yang dapat menentukan *centroid* awal yang tepat untuk memperbaiki kinerja algoritma WDM K-Modes.

Salah satu algoritma yang dapat melakukan perbaikan penentuan *centroid* awal agar dapat menghasilkan *centroid* terbaik pada WDM K-Modes adalah *Evidence Accumulation* (EA) (Shri dan Shehroz, 2008). Algoritma *Evidence Accumulation* merupakan algoritma yang dikembangkan agar dapat menutupi kekurangan pada *K-Modes* yaitu terhadap penentuan *centroid* awal sehingga algoritma *Evidence Accumulation* mampu menentukan *centroid* awal yang tepat dan menghasilkan akurasi *cluster* yang lebih besar (Shri dan Shehroz, 2008).

Algoritma *Evidence Accumulation* bekerja sebelum dan setelah proses K-Modes. Awalnya, algoritma K-Modes diproses untuk memperoleh modus dari tiap iterasi. Kemudian *Evidence Accumulation* bekerja menyimpan modus di setiap iterasi K-Modes dan memilih yang paling beragam set modus yang dapat dimanfaatkan sebagai *centroid* awal untuk algoritma K-Modes.

Tabel 1.1 merupakan perbandingan antara kelebihan-kelebihan yang terdapat pada masing-masing algoritma.

Tabel 1.1 Perbandingan Kelebihan dan Kekurangan Algoritma

| Algoritma | Penentuan Pengukuran Jarak *Cluster* | Penentuan *Centroid* Awal |
|---|---|---|
| K-Modes (Z Huang, 1998) | – | – |
| WDM K-Modes (Aranganayagi dan Thangavel, 2009) | √ | – |
| *Evidence Accumulation* (Shri dan Shehroz, 2008) | – | √ |

Berdasarkan Tabel 1.1, maka terlihat bahwa algoritma WDM K-Modes dan *Evidence Accumulation* memiliki tugas yang berbeda, dimana WDM K-Modes menambahkan algoritma baru untuk penentuan pengukuran jarak *cluster* dan *Evidence Accumulation* menambahkan algoritma baru untuk penentuan *centroid* awal. Penelitian ini akan menggabungkan *Evidence Accumulation* ke dalam algoritma WDM K-Modes sebagai usaha perbaikan *centroid* awal sehingga dapat menghasilkan nilai *purity* yang lebih tinggi. Kemudian membandingkan K-Modes tradisional, WDM K-Modes dan WDM-EA K-Modes, *dataset* mana yang akan menghasilkan nilai *purity* yang lebih tinggi. Penelitian ini juga akan melakukan evaluasi dengan menghitung akurasi terhadap hasil *clustering* menggunakan metode *Purity Measure* (Z Huang, 1998). Sehingga implementasi dapat membentuk *cluster* dalam pengelompokan objek yang diaplikasikan dalam kasus dunia nyata.

## 1.2 Identifikasi Masalah

Berikut identifikasi masalah dari penelitian ini :

1) Pengukuran jarak pada K-Modes tidak mempertimbangkan hubungan kesamaan yang ada pada nilai atribut sehingga menghasilkan kesamaan *intra cluster* yang lemah (Z Huang, 1998)
2) Penentuan centroid awal sangat penting bagi algoritma K-Modes, tetapi tidak ada ketentuan yang berlaku untuk menentukan *centroid* awal yang mana yang akan dibentuk (Shri dan Shehroz, 2008)

Dari permasalahan tersebut, munculah sebuah tantangan yaitu bagaimana cara atau langkah untuk menentukan pengukuran jarak dan *centroid* awal yang tepat pada K-Modes sehingga dapat menghasilkan kualitas *cluster* yang baik, sehingga muncul ide untuk menggabungkan dua algoritma yang dapat menutupi kekurangan dari K-Modes yaitu algorima WDM K-Modes dan *Evidence Accumulation*.

## 1.3 Tujuan Penelitian

Tujuan dari penelitian ini adalah sebagai berikut :

1) Menggabungkan algoritma WDM dan *Evidence Accumulation* untuk menghasilkan *cluster* yang memiliki nilai *Purity* terbaik
2) Mengetahui karakteristik gabungan algoritma WDM dan *Evidence Accumulation*

## 1.4 Manfaat Penelitian

Manfaat yang dapat diperoleh dalam penelitian ini adalah:

1) Pada sisi keilmuan, kontribusi utama yang dihasilkan dari penelitian ini adalah dihasilkannya akurasi *clustering* dengan cara mencari nilai *purity*

ketika implementasi metode *Evidence Accumulation* pada algoritma WDM K-Modes untuk perbaikan nilai *purity*.

2) Bagi kegiatan penelitian *clustering*, penelitian ini memberikan kontribusi sebagai pembuka topik-topik penelitian berikutnya. Penelitian-penelitian tersebut meliputi pengembangan konsep *data mining*, *clustering*, pengukuran jarak *cluster* dan pengembangan algoritma K-Modes, serta implementasi algoritma yang diusulkan pada studi kasus di dunia nyata.

## 1.5 Lingkup Masalah

Lingkup masalah dari penelitian ini adalah sebagai berikut :

1) Perbaikan proses *clustering* berfokus pada jarak *cluster* dan *centroid* awal pada algoritma K-Modes.
2) Algoritma yang diuji adalah algoritma *Weighted Dissimilarity Measure* K-Modes dan *Evidence Accumulation* K-Modes.
3) Penelitian berfokus pada peningkatan nilai *purity* hasil *clustering*.
4) *Dataset* yang digunakan dalam penelitian ini sama dengan *dataset* yang digunakan pada penelitian lain yang mempunyai atribut kategori, yaitu *Soybean Small, Zoo, Balance Scale dan Car Evaluation*.

## 1.6 Sistematika Penulisan

Laporan penelitian ini disusun dengan sistematika sebagai berikut.

**Bab 1 Pendahuluan**

Bab 1 menjelaskan latar belakang penelitian, identifikasi masalah, tujuan dan manfaat penelitian, batasan masalah, dan sistematika penelitian.

**Bab 2 Landasan Teori**

Bab 2 membahas pengertian algoritma K-Modes, konsep dari metode WDM K-Modes dan *Evidence Accumulation* serta masalah-masalah pada algoritma K-Modes. Pada bagian akhir bab ini dibahas hasil penelitian lain yang terkait dengan penelitian ini.

**Bab 3 Metode Penelitian**

Bab 3 membahas tahap desain perbaikan algoritma WDM K-Modes menggunakan *Evidence Accumulation* serta realisasinya pada dataset. Pembahasan terdiri dari karakteristik *dataset,* proses penggabungan algoritma, analisis data dan bagaimana cara pengujiannya.

**Bab 4 Hasil Pengujian dan Pembahasan**

Bab 4 membahas hasil pengujian dari hasil *clustering* yang telah terbentuk. Kemudian mengimplementasikan pada beberapa *dataset* dan menghasilkan nilai akurasi (*purity*) pada algoritma WDM K-Modes, *Evidence Accumulation* dan WDM-EA K-Modes.

**Bab 5 Penutup**

Bab 5 menyimpulkan hasil-hasil yang diperoleh dari penelitian ini dan tindak lanjut untuk penelitian tahap berikutnya.